

# Geometric Aspects of Biological Sequence Comparison

Aleksandar Stojmirović and Yi-Kuo Yu

National Center for Biotechnology Information,  
National Library of Medicine, National Institutes of Health,  
Bethesda, MD 20894, United States

February 2, 2008

## Abstract

We introduce a geometric framework suitable for studying the relationships among biological sequences. In contrast to previous works, our formulation allows asymmetric distances (quasi-metrics), originating from uneven weighting of strings, which may induce non-trivial partial orders on sets of biosequences. The distances considered are more general than traditional generalized string edit distances. In particular, our framework enables non-trivial conversion between sequence similarities, both local and global, and distances. Our constructions apply to a wide class of scoring schemes and require much less restrictive gap penalties than the ones regularly used. Numerous examples are provided to illustrate the concepts introduced and their potential applications.

## 1 Introduction

Biological macromolecules such as DNA, RNA and proteins play an essential role in all living organisms. Structurally, they are all chains of residues belonging to a small set of basic molecules and the functional characteristics of each macromolecule are determined by the order and composition of

its components. It is therefore not surprising that comparison and alignment of biological sequences is one of the most important contributions from computational biology to modern biosciences.

Typical approaches to biosequence comparison are either distance- [67, 82] or similarity-based [55, 69]. The distance-based approaches minimize the cost, while those based on similarity maximize the likelihood of transformation of one sequence into another. In both cases the comparison scores for sequences are obtained by extension from scores over alphabets of basic molecules. The algorithms for computation of alignments are based on the dynamic programming technique [4]. Similarity-based methods became widely accepted because the Smith-Waterman algorithm [69] allows computation of local alignments, involving only parts of sequences to be compared. Local alignments are highly appropriate in biological context because elements of structure and function are usually restricted to discrete regions of biosequences and hence strong similarity of fragments of two sequences need not extend to similarity of full sequences. Most distance methods have been global in nature and could not be easily adapted for local comparison.

A downside of using local similarities for sequence comparison is that, while their statistics can be characterized [38, 42], no constraints, apart from algorithmic ones, are placed on the form that similarity measures can take. Under such conditions, sets of biosequences with similarity measures cannot be identified with mathematical structures such as metric or normed spaces, which are a natural framework for many computational techniques such as clustering [83] and indexing for similarity search [30]. In contrast, distance measures on sequences naturally correspond to metrics under some mild restrictions.

While the duality between global similarities and distances has been recognized very early [70], it was only recently established independently by Stojmirović [73] and by Spiro and Macura [71] that it is possible to transform local sequence similarity scores derived from many popular scoring functions on building blocks of DNA and proteins into distances satisfying the triangle inequality. In the contexts in which they were presented, the results of the above two papers are almost equivalent, however, their perspectives are quite different. Spiro and Macura [71] assume symmetric similarity scores and consider the transformation which converts a similarity to a metric, while [73] converts similarity into a quasi-metric, a metric without the symmetry axiom. Quasi-metrics naturally correspond to partial orders and are therefore a natural framework for local similarities.

Unlike most existing literature entries, which are concerned with alignment algorithms, this paper aims to show a rigorous connection between similarities and distances that are metrics or quasi-metrics. Our main results are presented in a form that allows transfer to domains that are not necessarily related to classical string transformations and for that reason we use the framework of free semigroups. We define the  $\ell^p$ -type edit distance, which generalizes the regular edit distance and allows us to consider many more scoring functions on the amino acid alphabet that fail the requirements in [73] and [71]. Our results also allow for similarities and distances that are asymmetric. In order to have an accurate description of distances generated from similarities, we introduce a novel nomenclature.

Section 2 presents the basic definitions. Edit distances and global similarities are discussed in Sections 3 and 4, respectively. Our main result, Theorem 5.3 is presented in Section 5 and various kinds of local similarities are discussed as examples. Section 6 examines the applicability of our theory to the actual similarity measures used in contemporary computational biology, while Section 7 discusses some possible applications of our results and future directions. We chose to state many of the well-known results formally and to present many examples to enhance readability. The proofs of the established results are either omitted, or, when generalized in our new framework, relegated to Appendix A.

## 2 Preliminaries

### 2.1 Sequences and Free Semigroups

Recall that the *free monoid* on a nonempty set  $\Sigma$ , denoted  $\Sigma^*$ , is the monoid whose elements, called *words* or *strings*, are all finite sequences of zero or more elements from  $\Sigma$ , with the binary operation of concatenation. The unique sequence of zero letters (empty string), which we shall denote  $e$ , is the identity element. The *free semigroup* on  $\Sigma$ , denoted  $\Sigma^+$  is the subset of  $\Sigma^*$  containing all elements except the identity.

The length of a word  $w \in \Sigma^*$ , denoted  $|w|$ , is the number of occurrences of members of  $\Sigma$  in it. For  $w = \sigma_1\sigma_2\ldots\sigma_n$ , where  $\sigma_i \in \Sigma$ ,  $|w| = n$  and we set  $|e| = 0$ .

For two words  $u, v \in \Sigma^*$ ,  $u$  is a *factor* or *substring* of  $v$  if  $v = xuy$  for some  $x, y \in \Sigma^*$  and  $u$  is a *subsequence* or *subword* of  $v$  if  $v = w_1^*u_1^*w_2^*u_2^*\ldots w_n^*u_n^*w_{n+1}^*$ ,

where  $u = u_1^* u_2^* \dots u_n^*$ ,  $u_i^* \in \Sigma^*$  and  $w_i^* \in \Sigma^*$ . For any  $x \in \Sigma^*$ , we use  $\mathfrak{F}(x)$  to denote the set of all factors of  $x$ .

We call a semigroup (monoid)  $(X, \star)$  *free* if it is isomorphic to the free semigroup (monoid) on some set  $\Sigma$ . The unique set of elements of  $X$  mapping to  $\Sigma$  under the isomorphism is called the set of *free generators*.

**Example 2.1.** A DNA molecule can be represented as a word in the free semigroup generated by the four-letter nucleotide alphabet  $\Sigma = \{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}$ . An RNA molecule is a word in the free semigroup generated by the alphabet  $\Sigma = \{\mathbf{A}, \mathbf{U}, \mathbf{C}, \mathbf{G}\}$ . A protein can be thought of as a word in the free semigroup generated by the standard twenty amino acid alphabet.

**Example 2.2.** Let  $\Sigma$  be a set and denote by  $\mathcal{M}(\Sigma)$  the set of all finite measures supported on  $\Sigma$ . We will call the elements of the free monoid  $\mathcal{M}(\Sigma)^*$  *profiles* over  $\Sigma^*$ . Profiles arise as models of sets of structurally related biological sequences where  $\Sigma$  is the nucleotide or amino acid alphabet.

As a convention, for any word  $u \in \Sigma^*$ , the notation  $u = u_1 u_2 \dots u_n$ , where  $n = |u|$  shall mean that  $u_i \in \Sigma$  while the notation  $u = u_1^* u_2^* \dots u_m^*$  shall imply that  $u_i^* \in \Sigma^*$ . For all  $1 \leq k \leq |u|$  we shall use  $\bar{u}_k$  to denote the word  $u_1 u_2 \dots u_k$  and set  $\bar{u}_0 = e$ .

Let  $f : \Sigma \rightarrow \mathbb{R}$ . The *canonical homomorphic extension of  $f$  to the free monoid  $\Sigma^*$*  is a function  $\bar{f} : \Sigma^* \rightarrow \mathbb{R}$  such that  $\bar{f}(e) = 0$  and for all  $x \in \Sigma^+$ ,  $\bar{f}(x) = \sum_{i=1}^{|x|} f(x_i)$ .

## 2.2 Quasi-metrics

Quasi-metrics are asymmetric distance functions that generalize metrics and partial orders. With their associated structures, they belong to an area of active research in topology and theoretical computer science [43]. We now produce the standard definitions used in the remainder of this paper.

A *quasi-metric* on a set  $X$  is a mapping  $d : X \times X \rightarrow \mathbb{R}_+$  such that for all  $x, y, z \in X$ :

$$(i) \quad d(x, y) = d(y, x) = 0 \iff x = y, \text{ and}$$

$$(ii) \quad d(x, z) \leq d(x, y) + d(y, z).$$

The axiom (ii) is known as the *triangle inequality*. If in addition  $d$  is symmetric, that is  $d(x, y) = d(y, x)$  for all  $x, y \in X$ , then  $d$  is called a *metric*.

A pair  $(X, d)$ , where  $X$  is a set and  $d$  a (quasi-) metric, is called a (quasi-) metric space.

For a quasi-metric  $d$ , its *conjugate* (or *dual*) quasi-metric, denoted  $d^*$ , is defined on  $X \times X$  by  $d^*(x, y) = d(y, x)$ , and its *associated metric*, denoted  $d^s$ , by  $d^s(x, y) = \max\{d(x, y), d(y, x)\} = d(x, y) \vee d^*(x, y)$ . Another frequently used symmetrization of a quasi-metric is the ‘sum’ metric  $d^u$  defined by  $d^u(x, y) = d(x, y) + d(y, x)$ .

A (left) open ball of radius  $r > 0$  centered at  $x_0 \in X$  with respect to a quasi-metric  $d$  is the set  $\{x \in X : d(x_0, x) < r\}$ . The collection of all (left) open balls centered at any  $x \in X$  with any  $r > 0$  is a base for a topology on  $X$  induced by  $d$ . This topology is in general  $T_0$  but not necessarily  $T_1$ . For the purpose of this paper, we will call a quasi-metric  $d$  *separating* if the induced topology is  $T_1$ , that is, if  $d(x, y) = 0$  implies  $x = y$  for all  $x, y \in X$ . Every quasi-metric  $d$  also has its *associated partial order*, denoted  $\leq_d$ , defined by  $x \leq_d y \iff d(x, y) = 0$ .

A quasi-metric  $d$  is called a *weightable quasi-metric* [44] if there exists a function  $w : X \rightarrow \mathbb{R}_+$ , called the *weight function* or simply the *weight*, satisfying for every  $x, y \in X$

$$d(x, y) + w(x) = d(y, x) + w(y).$$

In this case we call  $d$  *weightable* by  $w$ . A quasi-metric  $d$  is *co-weightable* if its conjugate quasi-metric  $d^*$  is weightable. The weight function  $w$  by which  $d^*$  is weightable is called the *co-weight* of  $d$  and  $d$  is *co-weightable* by  $w$ .

A concept strongly related to weighted quasi-metrics is that of a *partial metric* [50]. A *partial metric* on a set  $X$  is a mapping  $p : X \times X \rightarrow \mathbb{R}_+$  such that for all  $x, y, z \in X$ :

- (i)  $p(x, y) \geq p(x, x)$ ;
- (ii)  $x = y \iff p(x, x) = p(y, y) = p(x, y)$ ;
- (iii)  $p(x, y) = p(y, x)$ ;
- (iv)  $p(x, z) \leq p(x, y) + p(y, z) - p(y, y)$ .

It has been shown [50] that there is a bijection between the partial metrics and generalized weighted quasi-metrics: the transformation  $d(x, y) = p(x, y) - p(x, x)$  produces a generalized weighted quasi-metric with weight function  $x \mapsto p(x, x)$  out of a partial metric while the  $p(x, y) = q(x, y) + w(x)$  produces a partial metric out of a generalized weighted quasi-metric.

### 3 Edit distance

Waterman, Smith and Beyer, in their 1976 paper [82], introduced a general form of the edit distance on sets of words, henceforth referred to as the WSB distance. It was constructed by defining a set of allowed weighted transformations between two strings and then minimizing the sum of weights of allowed operations transforming (in the sense of ordered composition) one word into another. They also proposed an algorithm to compute the WSB distance based on dynamic programming.

In this section, we present a recursive definition of edit distance on a free semigroup that generalizes that of Waterman, Smith and Beyer and describe some of its most important properties. The edit distance provides the conceptual and algorithmic foundation to both global and local similarities on free semigroups. Before producing the main definition, we formalize the concept of a *gap penalty*, which we will discuss in detail later in the text.

**Definition 3.1.** Let  $\Sigma$  be a set. A positive function  $\gamma : \Sigma^+ \rightarrow \mathbb{R}$  is called a *gap penalty* over  $\Sigma^+$  if for all  $u, v \in \Sigma^+$ ,

$$\gamma(u) + \gamma(v) \geq \gamma(uv). \quad (1)$$

We denote by  $\Gamma(\Sigma)$  the set of all gap penalties over  $\Sigma^+$ .

**Definition 3.2.** Let  $\Sigma$  be a set,  $d : \Sigma \times \Sigma \rightarrow \mathbb{R}$ , and  $\alpha$  and  $\beta$  be functions  $\Sigma^+ \rightarrow \mathbb{R}$  such that  $\alpha^p, \beta^p \in \Gamma(\Sigma)$ . Let  $x, y \in \Sigma^*$  and let  $m = |x|$  and  $n = |y|$ . Let  $1 \leq p < \infty$  and define the distance  $D : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}$  using the following recursion:

- (a)  $D(\bar{x}_0, \bar{y}_0) = D(e, e) = 0$ ,
- (b)  $D(e, \bar{y}_j) = \alpha(\bar{y}_j)$  for all  $1 \leq j \leq n$ ,
- (c)  $D(\bar{x}_i, e) = \beta(\bar{x}_i)$  for all  $1 \leq i \leq m$ , and
- (d) for all  $1 \leq i \leq m$  and  $1 \leq j \leq n$

$$D(\bar{x}_i, \bar{y}_j) = \left( \min \left\{ D^p(\bar{x}_{i-1}, \bar{y}_{j-1}) + d^p(x_i, y_j), \right. \right. \\ \left. \min_{1 \leq k \leq j} \{ D^p(\bar{x}_i, \bar{y}_{j-k}) + \alpha^p(y_{j-k+1} \dots y_j) \}, \right. \\ \left. \min_{1 \leq k \leq i} \{ D^p(\bar{x}_{i-k}, \bar{y}_j) + \beta^p(x_{i-k+1} \dots x_i) \} \right\} \right)^{1/p}.$$

The  $\ell^p$  edit distance between the sequences  $x$  and  $y$  (extending  $d$ ,  $\alpha$  and  $\beta$ ), is then given by  $D(x, y) = D(\bar{x}_m, \bar{y}_n)$ .

**Remark 3.3.** We have assumed that  $\alpha^p, \beta^p \in \Gamma(\Sigma)$  instead of just being positive functions in order to have  $D(e, x) = \alpha(x)$  and  $D(x, e) = \beta(x)$  for all  $x \in \Sigma^+$ . For a general positive function  $\alpha : \Sigma^+ \rightarrow \mathbb{R}$ , the function  $\gamma$ , given recursively for all  $x \in \Sigma^+$  by  $\gamma(x_1) = \alpha^p(x_1)$  and

$$\gamma(\bar{x}_i) = \min_{1 \leq k \leq i} \{ \gamma(\bar{x}_{i-k}) + \alpha^p(x_{i-k+1} \dots x_i) \}, \quad (2)$$

will belong to  $\Gamma(\Sigma)$  and therefore  $\gamma^{1/p}$  can be used in definition of  $D$  instead of  $\alpha$ .

**Remark 3.4.** Also note that the distance  $D$  as defined does not extend  $d$  from  $\Sigma$  in the strict sense, that is, it is not necessarily true that for all  $a, b \in \Sigma$ ,  $D(a, b) = d(a, b)$ . However, this statement does become correct if we additionally assume  $d^p(a, b) \leq \beta^p(a) + \alpha^p(b)$ .

**Remark 3.5.** The  $\ell^p$  edit distance between  $x$  and  $y$  can be computed using dynamic programming algorithm of Waterman, Smith and Beyer [82]. Let  $\mathbf{D}$  be an  $(m+1) \times (n+1)$  matrix with rows and columns indexed from 0 such that  $\mathbf{D}_{0,0} = 0$  and for all  $i = 1, 2 \dots m$  and  $j = 1, 2 \dots n$ ,  $\mathbf{D}_{i,0} = \beta(\bar{x}_i)$ ,  $\mathbf{D}_{0,j} = \alpha(\bar{y}_j)$ , and

$$\mathbf{D}_{i,j} = \min \left\{ \begin{aligned} &\mathbf{D}_{i-1,j-1} + d^p(x_i, y_j), \\ &\min_{1 \leq k \leq j} \{ \mathbf{D}_{i,j-k} + \alpha^p(y_{j-k+1} \dots y_j) \}, \\ &\min_{1 \leq k \leq i} \{ \mathbf{D}_{i-k,j} + \beta^p(x_{i-k+1} \dots x_i) \} \end{aligned} \right\}. \quad (3)$$

Then, we have  $D(x, y) = (\mathbf{D}_{m,n})^{1/p}$ . The original WSB distance is obtained when  $p = 1$ .

### 3.1 Alignments

From the recursive definition, it follows that the  $\ell^p$  edit distance  $D(x, y)$  can be decomposed as the  $\ell^p$  sum of the distances of non-overlapping factors of  $x$  and  $y$ . This decomposition provides an optimal *alignment* between  $x$  and  $y$ .

**Definition 3.6** ([68]). Let  $x, y \in \Sigma^*$ . An *alignment* between  $x$  and  $y$  is a finite sequence of pairs  $\langle (x_k^*, y_k^*) \rangle_{k=1}^K$ , where  $x = x_1^* x_2^* \dots x_K^*$ ,  $y = y_1^* y_2^* \dots y_K^*$  and for each  $1 \leq k \leq K$  either

- (a)  $x_k^* = x_i$  and  $y_k^* = y_j$  for some  $i, j$ , or
- (b)  $x_k^* \in \mathfrak{F}(x)$ ,  $x_k^* \neq e$  and  $y_k^* = e$ , or
- (c)  $x_k^* = e$ ,  $y_k^* \in \mathfrak{F}(y)$  and  $y_k^* \neq e$ .

We will use  $\mathcal{A}(x, y)$  to denote the set of all alignments of  $x$  and  $y$ .

Each pair  $(x_k^*, y_k^*)$  corresponds to an *edit operation* that transforms  $x_k^*$  into  $y_k^*$ . Pairs of the form  $(a, b)$ ,  $(x, e)$  and  $(e, y)$  where  $a, b \in \Sigma$  and  $x, y \in \Sigma^+$  represent a *substitution* of the letter  $a$  for the letter  $b$ , *deletion* of the word  $x$  and *insertion* of the word  $y$ , respectively. Insertions and deletions are collectively called *indels*.

Every transformation  $(x_k^*, y_k^*)$  can be given a weight or a cost equal to  $D(x_k^*, y_k^*)$ , with the weight of an alignment  $\langle (x_k^*, y_k^*) \rangle_{k=1}^K$  being equal to the  $\ell^p$  sum of the weights of the individual transformations. The distance  $d$  on  $\Sigma$  provides *substitution costs*, while the values of  $\alpha$  and  $\beta$ , give the costs of indels. Thus, the edit distance between  $x$  and  $y$  can be described as the minimum weighted cost (in the  $\ell^p$  sense) of transforming the sequence  $x$  into  $y$  using substitutions and indels as edit operations. This provides an alternative characterization of edit distance, which was long known for the  $\ell^1$  case [68] and which we state here in general form without proof as Lemma 3.7 below.

**Lemma 3.7.** *Let  $\Sigma$  be a set,  $d : \Sigma \times \Sigma \rightarrow \mathbb{R}$ , and  $\alpha, \beta : \Sigma^+ \rightarrow \mathbb{R}_+$ . Suppose  $D$  is an  $\ell^p$  edit distance on  $\Sigma^*$  with respect to  $d$ ,  $\alpha$  and  $\beta$ . Then, for all  $x, y \in \Sigma^*$*

$$D(x, y) = \min \left\{ \left( \sum_{k=1}^K D^p(x_k^*, y_k^*) \right)^{1/p} \mid \langle (x_k^*, y_k^*) \rangle_{k=1}^K \in \mathcal{A}(x, y) \right\}. \quad (4)$$

□

### 3.2 Edit distances as quasi-metrics

We now proceed to state the conditions for an  $\ell^p$  edit distance to be a quasi-metric. For simplicity we restrict ourselves to edit distances with gap



penalties that are increasing and depend solely on fragment composition and length, while more general gap penalties are considered in Appendix A.1.

**Definition 3.8.** Let  $\Sigma$  be a set. We call a function  $\gamma : \Sigma^* \rightarrow \mathbb{R}$  *increasing* if for all  $u, v, x \in \Sigma^*$ ,

$$\gamma(uxv) \geq \gamma(uv). \quad (5)$$

**Definition 3.9.** Let  $\Sigma$  be a set. A function  $\gamma \in \Gamma(\Sigma)$  is called a *composition-length gap penalty* on  $\Sigma^+$  if it is increasing and has a form

$$\gamma(z) = \sum_i \phi(z_i) + \psi(|z|) \quad (6)$$

for all  $z \in \Sigma^+$ , where  $\phi$  is a map  $\Sigma \rightarrow \mathbb{R}$  and  $\psi$  is a function  $\mathbb{N} \rightarrow \mathbb{R}$ . We denote by  $\Gamma_{CL}(\Sigma)$  the set of all composition-length gap penalties on  $\Sigma^+$ .

Composition-length gap penalties have a component solely dependent on the length of the inserted or deleted word and a composition-dependent component. Current applications of edit distances in computational biology (see for example [25]) mainly use gap penalties that are the same for insertions and deletions and depend solely on the fragment length, thus satisfying our definition of composition-length gap penalties with  $\phi = 0$ . We chose the above definition in order to include all such cases and to provide simple but sufficiently general gap penalties for consideration of global and local similarities. The requirement for composition-length gap penalties to be increasing is included because it is a necessary condition for applications of our main Theorem 5.3.

The most widely used length-dependent gap penalty functions are *linear*, of the form  $\psi(k) = \mu k$ , and *affine*, of the form  $\psi(k) = \mu + \nu k$ , where  $\mu, \nu$  are constants. The main advantage of affine gap penalties is that the dynamic programming algorithm for computation of distances in this case can be modified to run in  $O(nm)$  average and worst case time, where  $m = |x|$  and  $n = |y|$  [21], as opposed to  $O(m^2n + mn^2)$  for the most general WSB algorithm [82]. Gap penalties of the form  $\psi(k) = \mu + \nu \log(k)$  have also been considered [81]. Note that the algorithmic complexity of the WSB algorithm for distances using composition-length gap penalties depends mainly on the form of  $\psi$  since the composition-dependent component is linear.

**Theorem 3.10.** Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . Suppose  $d$  is a separating quasi-metric on  $\Sigma$  and  $\gamma, \delta \in \Gamma_{CL}(\Sigma)$  such that for all  $a, b \in \Sigma$ ,

$$\gamma(b) - \gamma(a) \leq d^p(a, b) \quad (7)$$

and

$$\delta(a) - \delta(b) \leq d^p(a, b). \quad (8)$$

Let  $\alpha = \gamma^{1/p}$  and  $\beta = \delta^{1/p}$ . Then, the  $\ell^p$  edit distance  $D$ , extending  $d, \alpha$  and  $\beta$ , is a separating quasi-metric on  $\Sigma^*$ .  $\square$

Theorem 3.10 is a generalization of similar theorems for  $p = 1$  proven by Waterman *et al.* [82] for constant substitution costs and gap penalties depending on fragment length, and by Spiro and Macura [71] in a more general setting. We state and prove a version with fewer restriction on gap penalties as Theorem A.1 in Appendix A.1.

**Remark 3.11.** According to [64], a quasi-metric  $d$  defined on a semigroup  $(X, \star)$  is called *invariant* with respect to  $\star$  if for all  $x, y, z \in X$ ,

$$d(x \star z, y \star z) \leq d(x, y) \quad \text{and} \quad d(z \star x, z \star y) \leq d(x, y). \quad (9)$$

It is apparent from the definition that the edit distance  $D$  on the free semigroup  $\Sigma^*$ , which satisfies Theorem 3.10, is invariant with respect to the string concatenation.

Since our  $\ell^p$  edit distances depend on several parameters, we introduce a nomenclature to make this explicit.

**Definition 3.12.** Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . Suppose  $D$  is an  $\ell^p$  edit distance extending a quasi-metric  $d$  on  $\Sigma$  and gap penalties  $\alpha, \beta$  such that  $\alpha^p, \beta^p \in \Gamma_{\text{CL}}(\Sigma)$ . We will write  $D = \text{EQ}^p(d, \alpha, \beta)$  if  $D$  is a quasi-metric and  $D = \text{EM}^p(d, \alpha)$  if  $D$  is a metric (it is necessary that  $\alpha = \beta$  if  $D$  is a metric).

Most (if not all) instances of edit distances in computer science, computational biology and pure mathematics involve the  $\ell^1$  edit distances. Below, we outline some of the well-known examples.

**Example 3.13.** The *Levenshtein metric* [46] (the original ‘string edit distance’) is the smallest number of permitted edit operations (substitutions and indels) required to transform one string into another. In our nomenclature, for a set of letters  $\Sigma$ , the Levenshtein distance is realized as  $\text{EM}^1(d, \alpha)$  where  $\alpha(u) = |u|$  for all  $u \in \Sigma^+$  and  $d$  is the *discrete metric*, that is, for all  $a, b \in \Sigma$

$$d(a, b) = \begin{cases} 0 & \text{if } a = b, \\ 1 & \text{if } a \neq b. \end{cases} \quad (10)$$

**Example 3.14.** The Sellers distance, introduced by Sellers in 1974 [67], is a metric obtained by extension of a metric  $d$  on the set  $\Sigma_{\dagger} = \Sigma \cup \{e\}$ , the set of generators plus the identity element, to the free monoid  $\Sigma^*$ . It is realized as  $\mathbf{EM}^1(d, \alpha)$  where  $\alpha(u) = \sum_i d(u_i, e)$  for all  $u \in \Sigma^+$ .

This construction has long been known in the theory of topological groups [59] as the Graev metric [22, 23] on the free group  $F(\Sigma)$ . Recall that  $F(\Sigma)$  consists of all sequences of letters from the generating set  $\Sigma$  and their inverses; in other words,  $F(\Sigma) = Y^*$ , where  $Y = \Sigma \cup \Sigma^{-1}$  and  $\Sigma^{-1}$  is the set consisting of inverses of elements of  $\Sigma$ . Let  $\rho$  be a metric on the set  $Y_{\dagger} = Y \cup \{e\}$ . The Graev metric  $\bar{\rho}$  is then a maximal invariant metric on  $F(\Sigma)$  such that  $\bar{\rho}$  restricted to the set  $Y_{\dagger}$  is equivalent to  $\rho$ . Note that the notion of invariance in this context is slightly different than the definition of an invariant quasi-metric on a semigroup from Remark 3.11 above: a metric  $\rho$  on a group  $(X, \star)$  is called *invariant* with respect to  $\star$  if for all  $x, y, z \in X$ ,

$$\rho(x \star z, y \star z) = \rho(z \star x, z \star y) = \rho(x, y). \quad (11)$$

The maximality of the Sellers-Graev metric can also be observed in the context of the free monoid  $\Sigma^*$  using the following argument. Let  $D = \mathbf{EM}^1(d, \alpha)$  where  $d$  is a on  $\Sigma$  and  $\alpha$  is a gap penalty. Define a metric  $d_{\dagger}$  on  $\Sigma_{\dagger}$  by

$$d_{\dagger}(a, b) = \begin{cases} D(a, b) & \text{if } a, b \in \Sigma, \\ \alpha(a) & \text{if } b = e, \\ \alpha(b) & \text{if } a = e. \end{cases} \quad (12)$$

It is clear that  $D$  extends  $d_{\dagger}$  from  $\Sigma_{\dagger}$  to  $\Sigma^*$ . However, for every  $x \in \Sigma^*$ ,

$$D(x, e) \leq \left( \sum_i \alpha^p(x_i) \right)^{1/p} \leq \sum_i \alpha(x_i)$$

and hence every edit distance extending  $d_{\dagger}$  to  $\Sigma^*$  will be smaller than the Sellers-Graev distance.

**Example 3.15.** Let  $\Sigma$  be a set and for  $u, v \in \Sigma^*$  denote by  $\mathbf{LCS}(u, v)$  the *longest common subsequence* of  $u$  and  $v$ . Define

$$\rho(u, v) = |u| + |v| - 2|\mathbf{LCS}(u, v)|.$$

It can be easily shown that  $\rho$  is a metric on  $\Sigma^*$  and that  $\rho$  can be realized as  $\mathbf{EM}^1(d, \alpha)$  where  $\alpha(u) = |u|$  for all  $u \in \Sigma^+$  and  $d(a, b) = 2$  for all  $a, b \in \Sigma$

such that  $a \neq b$  (cf. [25], pp. 246). Since  $d(a, b) \geq \alpha(a) + \alpha(b)$ , the optimal alignment can be expressed solely in terms of insertions and deletions. The longest common subsequence metric provides a special case of the Sellers-Graev metric.

### 3.3 Alignment decomposition

Recall that Lemma 3.7 indicates that the total  $\ell^p$  edit distance  $D$  between two words  $x$  and  $y$  can be optimally decomposed as an  $\ell^p$  sum of the distances between constituent factors of  $x$  and  $y$ . Lemma 3.17 below shows that, if the gap penalties are increasing, an arbitrary choice of a factor  $y'$  of  $y$  decomposes the edit distance between  $x$  and  $y$  into  $\ell^p$  sum of the edit distances between fragments of  $x$  and  $y$ . In this case, all of  $x$  is used up while some parts of  $y$  could be ‘lost’ (Figure 1). A similar splitting can also be achieved with a choice of a fragment of  $x$ . We call this property *arbitrary decomposability*.

**Definition 3.16.** Let  $\Sigma$  be a set, let  $\rho : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}$  be a distance function on the free monoid  $\Sigma^*$  and let  $1 \leq p < \infty$ . We say that  $\rho$  is *arbitrarily decomposable of order  $p$*  if for all  $x, y \in \Sigma^*$ ,

- (i) For every  $y' \in \mathfrak{F}(y)$  there exist  $x', x_1^*, x_2^* \in \mathfrak{F}(x)$  such that  $x = x_1^* x' x_2^*$  and  $y_1^*, y_2^*, u, v \in \mathfrak{F}(y)$  such that  $y = y_1^* u y' v y_2^*$  and

$$\rho(x, y) \geq \left( \rho^p(x_1^*, y_1^*) + \rho^p(x', y') + \rho^p(x_2^*, y_2^*) \right)^{1/p}; \quad (\text{A1})$$

- (ii) For every  $x' \in \mathfrak{F}(x)$  there exist  $y', y_1^*, y_2^* \in \mathfrak{F}(y)$  such that  $y = y_1^* y' y_2^*$  and  $x_1^*, x_2^*, u, v \in \mathfrak{F}(x)$  such that  $x = x_1^* u x' v x_2^*$  and

$$\rho(x, y) \geq \left( \rho^p(x_1^*, y_1^*) + \rho^p(x', y') + \rho^p(x_2^*, y_2^*) \right)^{1/p}. \quad (\text{A2})$$

Note that if the distance function  $\rho$  is symmetric, the two properties above collapse into a single one.

**Lemma 3.17.** Let  $\Sigma$  be a set and let  $d : \Sigma \times \Sigma \rightarrow \mathbb{R}$ . Suppose that  $\alpha$  and  $\beta$  are increasing functions  $\Sigma^+ \rightarrow \mathbb{R}$  such that  $\alpha^p, \beta^p \in \Gamma(\Sigma)$  and  $D$  is an  $\ell^p$  edit distance on  $\Sigma^*$  extending  $d$ ,  $\alpha$  and  $\beta$ . Then,  $D$  is arbitrarily decomposable of order  $p$ .

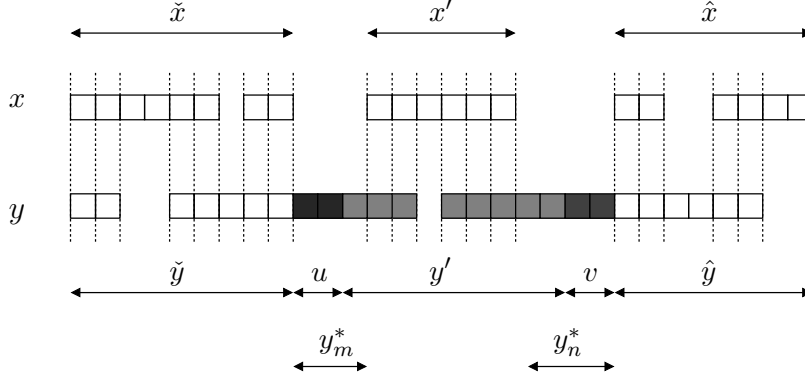


Figure 1: Arbitrary decomposability (part A1) of an alignment. A choice of  $y'$  induces a decomposition of both  $x$  and  $y$  such that  $x = \tilde{x}x'\hat{x}$ ,  $y = \tilde{y}uy'v\hat{y}$  and  $\rho(x, y) \geq (\rho^p(\tilde{x}, \tilde{y}) + \rho^p(x', y') + \rho^p(\hat{x}, \hat{y}))^{1/p}$ . Dashed lines indicate the boundaries of edit operations. The fragments  $u$  and  $v$  of  $y$  are ‘lost’: they do not contribute to decomposition.

*Proof.* We will prove only the first part of the definition of arbitrary decomposability because the second follows by the same argument. Let  $x, y \in \Sigma^*$  and let  $y' \in \mathfrak{F}(y)$ . By Lemma 3.7, the distance  $D(x, y)$  can be written as

$$D(x, y) = \left( \sum_{k=1}^K D^p(x_k^*, y_k^*) \right)^{1/p},$$

where  $x = x_1^* x_2^* \dots x_K^*$ ,  $y = y_1^* y_2^* \dots y_K^*$ . Let  $1 \leq m \leq n \leq K$  be such that  $y_m^* \neq e$ ,  $y_n^* \neq e$ ,  $y' \in \mathfrak{F}(y_m^* \dots y_n^*)$  and  $y_{m+1}^* \dots y_{n-1}^* \in \mathfrak{F}(y')$  (i.e.  $y_m^* \dots y_n^*$  is the smallest factor of  $y$  having  $y'$  as a factor – see Figure 1). Then, the fragments  $y_m^*$  and  $y_n^*$  contain parts of  $y'$ . (Note that  $y'$  always coincides with  $y_m^* \dots y_n^*$  if the gap penalties depend only on composition.)

Consider the fragment  $y_m^*$ . According to Lemma 3.7,  $y_m^*$  can be either a letter ( $y_m^* \in \Sigma$ ) or a fragment ( $y_m^* \in \Sigma^*$ ), since the possibility of  $y_m^* = e$  was explicitly excluded. If  $y_m^* \in \Sigma$ , let  $u = e$  and  $u' = y_m^*$  so that  $D(x_m^*, y_m^*) = D(x_m^*, u')$ . On the other hand, if  $y_m^* \notin \Sigma$ , then by Lemma 3.7  $x_m^* = e$ . Let  $u, u' \in \Sigma^*$  be fragments of  $y_m^*$  such that  $y_m^* = uu'$  and  $u'_1 = y'_1$  (i.e. we split  $y_m^*$  into a part not overlapping with  $y'$  and a part overlapping with it). It is possible that  $u = e$  but we always have  $u' \in \Sigma^+$  by construction. By our

assumption about increasing gap penalty, it follows that

$$D(x_m^*, y_m^*) = D(e, uu') = \alpha(uu') \geq \alpha(u) = D(x_m^*, u'). \quad (13)$$

In a similar way, the fragment  $y_n^*$  can be expressed as  $y_n^* = v'v$  where  $y'_{|y'|} = v'_{|v'|}$  (i.e.  $v'$  contains the end of  $y'$ ) and

$$D(x_n^*, y_n^*) = D(e, v'v) = \alpha(v'v) \geq \alpha(v) = D(x_n^*, v'). \quad (14)$$

Now, let  $\tilde{x} = x_1^* \dots x_{m-1}^*$ ,  $x' = x_m^* \dots x_n^*$  and  $\hat{x} = x_{n+1}^* \dots x_K^*$ . Let  $\tilde{y} = y_1^* \dots y_{m-1}^*$  and  $\hat{y} = y_{n+1}^* \dots y_K^*$ . Then,  $x = \tilde{x}x'\hat{x}$ ,  $y = \tilde{y}y'v\hat{y}$  and

$$\begin{aligned} D(x, y) &= \left( \sum_{k=1}^K D^p(x_k^*, y_k^*) \right)^{1/p} \\ &= \left( D^p(\tilde{x}, \tilde{y}) + D^p(x_m^*, uu') + \sum_{k=m+1}^{n-1} D^p(x_k^*, y_k^*) + D^p(x_n^*, v'v) + D^p(\hat{x}, \hat{y}) \right)^{1/p} \\ &\geq \left( D^p(\tilde{x}, \tilde{y}) + D^p(x_m^*, u') + \sum_{k=m+1}^{n-1} D^p(x_k^*, y_k^*) + D^p(x_n^*, v') + D^p(\hat{x}, \hat{y}) \right)^{1/p} \\ &\geq \left( D^p(\tilde{x}, \tilde{y}) + D^p(x', y') + D^p(\hat{x}, \hat{y}) \right)^{1/p}, \end{aligned}$$

since  $(x_m^*, u')(x_{m+1}^*, y_{m+1}^*) \dots (x_{n-1}^*, y_{n-1}^*)(x_n^*, v')$  is an alignment of  $x'$  and  $y'$  and hence the  $\ell^p$  sum of distances over it is greater than  $D^p(x', y')$  by Lemma 3.7.  $\square$

Therefore, any  $\ell^p$  edit distance with composition-length gap penalties is arbitrarily decomposable of order  $p$ . However, there exist arbitrarily decomposable distances that are not  $\ell^p$  edit distances.

**Example 3.18.** Let  $\Sigma$  be a finite set and let  $d$  be a metric on  $\Sigma$ . For any  $n \in \mathbb{N}$ , the *generalized Hamming distance*  $d_n$  on  $\Sigma^n$  is given for all  $x, y \in \Sigma^n$  by

$$d_n(x, y) = \sum_{i=1}^n d(x_i, y_i). \quad (15)$$

It can be easily shown that  $d_n$  is a metric. The generalized Hamming distance is a natural generalization of the Hamming distance [26] where the distance  $d$  on  $\Sigma$  is the discrete metric.

Let  $f : \Sigma \rightarrow \mathbb{R}$  be a function such that for all  $a, b \in \Sigma$ ,

$$|f(a) - f(b)| \leq d(a, b) \leq f(a) + f(b). \quad (16)$$

It immediately follows that for every  $n \in \mathbb{N}$  and for all  $x, y \in \Sigma^n$ ,

$$|\bar{f}(x) - \bar{f}(y)| \leq d_n(x, y) \leq \bar{f}(x) + \bar{f}(y). \quad (17)$$

Define the distance  $\rho : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}$  by extending  $d_n$  and  $f$  so that for all  $x, y \in \Sigma^*$ ,

$$\rho(x, y) = \begin{cases} d_n(x, y) & \text{if } |x| = |y| = n, \\ \bar{f}(x) + \bar{f}(y) & \text{if } |x| \neq |y|. \end{cases} \quad (18)$$

Using (17), it is easy to show that  $\rho$  is a metric on  $\Sigma^*$ . Furthermore,  $\rho$  is arbitrarily decomposable (of order 1). Indeed, consider  $x, y \in \Sigma^*$  and  $y' \in \mathfrak{F}(y)$ . If  $|x| = |y|$ , one immediately obtains the required decomposition using the form of the generalized Hamming distance. On the other hand, if  $|x| \neq |y|$ , we have

$$\rho(x, y) = \bar{f}(x) + \bar{f}(y) \geq \rho(x, e) + \rho(e, y') \quad (19)$$

leading to the decomposition where  $x' = e$  and  $u$  and  $v$  take all of  $y$  apart from  $y'$ .

The metric  $\rho$  (generalized to  $\ell^p$  form) can be interpreted as an ‘ungapped’ version of edit distances. Here substitutions are allowed only between sequences of equal length and the function  $\bar{f}$  plays a role of gap penalty so that the only way to transform sequences of unequal length is through a full deletion followed by insertion.

## 4 Global Similarity

A more common approach to sequence comparison is to maximize similarities instead of minimizing distances. In this case a similarity measure on  $\Sigma$  and gap penalties are used to define the similarity between two sequences in  $\Sigma^*$  using the Needleman-Wunsch [55] or Smith-Waterman [69] dynamic programming algorithm, which are very similar to the algorithm for computation of edit distances described above. As in the case of  $\ell^p$  edit distances above, we define sequence similarities using a recursive definition.

**Definition 4.1.** Let  $\Sigma$  be a set,  $s : \Sigma \times \Sigma \rightarrow \mathbb{R}$ , and let  $\gamma, \delta \in \Gamma(\Sigma)$ . For any  $x, y \in \Sigma^*$  where  $m = |x|$  and  $n = |y|$ , define the *global (Needleman-Wunsch) similarity*  $S : (x, y) \mapsto \mathbb{R}$  using the following recursion:

- (a)  $S(\bar{x}_0, \bar{y}_0) = S(e, e) = 0$ ,
- (b)  $S(e, \bar{y}_j) = -\gamma(\bar{y}_j) = \text{for all } 1 \leq j \leq n$ ,
- (c)  $S(\bar{x}_i, e) = -\delta(\bar{x}_i) = \text{for all } 1 \leq i \leq m$ , and
- (d) for all  $1 \leq i \leq m$  and  $1 \leq j \leq n$

$$S(\bar{x}_i, \bar{y}_j) = \max \left\{ \begin{aligned} &S(\bar{x}_{i-1}, \bar{y}_{j-1}) + s(x_i, y_j), \\ &\max_{1 \leq k \leq j} \{S(\bar{x}_i, \bar{y}_{j-k}) - \gamma(y_{j-k+1} \dots y_j)\}, \\ &\max_{1 \leq k \leq i} \{S(\bar{x}_{i-k}, \bar{y}_j) - \delta(x_{i-k+1} \dots x_i)\} \end{aligned} \right\}. \quad (20)$$

The global similarity between the sequences  $x$  and  $y$  (extending  $s$ ,  $\gamma$  and  $\delta$ ), is defined by  $S(x, y) = S(\bar{x}_m, \bar{y}_n)$ .

The algorithm used to compute  $\ell^1$  edit distance (Remark 3.5) can also be used for computation of similarities by setting  $d = -s$ ,  $\alpha = \gamma$  and  $\beta = \delta$ , computing  $D$  for  $p = 1$  and then taking  $S = -D$ . The running time of the dynamic programming algorithm depends on the properties of gap penalties, as discussed in the previous section. Note that the gap penalty functions are positive in the case of both distances and similarities, being added in the former case and subtracted in the latter. It is also possible to express global similarity as a sum of similarities over alignments, as is done for edit distance in Lemma 3.7.

**Example 4.2.** It is well known [25] that the longest common subsequence problem described in Example 3.15 can be approached using similarities rather than distances. Let  $\Sigma$  be a set and let  $s$  be a scoring function on  $\Sigma$  such that  $s(a, b) = 0$  if  $a \neq b$  and  $s(a, a) = 1$ . Let  $\gamma(x) = \delta(x) = 0$  for all  $x \in \Sigma^+$ . It is easy to confirm that for  $x, y \in \Sigma^*$ ,  $S(x, y) = |\text{LCS}(x, y)|$ .

Relations between global similarities and  $\ell^1$  edit distances were explored early on [70, 68].



**Theorem 4.3** ([70, 68]). *Let  $S$  be the global similarity with respect to  $s, \gamma$  and  $\delta$  such that for all  $x \in \Sigma^+$ ,  $\gamma(x) = \delta(x) = \psi(|x|)$ , where  $\psi$  is a positive function. Consider the  $\ell^1$  edit distance  $D$ , extending  $d : \Sigma \times \Sigma \rightarrow \mathbb{R}$  and the gap penalties  $\alpha$  and  $\beta$  and let  $s_M = \max\{s(a, b) \mid a, b \in \Sigma\}$ . Suppose for all  $a, b \in \Sigma$*

$$d(a, b) = s_M - s(a, b), \quad (21)$$

*and for all  $x \in \Sigma^+$ ,*

$$\alpha(x) = \beta(x) = \frac{s_M |x|}{2} + \psi(|x|). \quad (22)$$

*Then,  $S$  and  $D$  will induce equivalent sets of optimal alignments and for all  $x, y \in \Sigma^*$ ,*

$$D(x, y) = s_M \frac{|x| + |y|}{2} - S(x, y). \quad (23)$$

□

The distance function obtained by taking a constant minus similarity is not guaranteed to satisfy any of the axioms for a metric or a quasi-metric: one problem is that the self-similarity  $S(x, x)$  for any  $x \in \Sigma^*$  is not necessarily a constant. However, under some more restrictive but frequently valid assumptions, it is possible to transform similarities into metrics or quasi-metrics. We establish the results that have interesting biological interpretations and provide the foundation for considering transformation of local similarities, discussed in Section 5, to quasi-metrics.

**Definition 4.4.** Let  $X$  be a set and let  $s$  be a (similarity) map  $X \times X \rightarrow \mathbb{R}$ . We call  $s$  a *sane* scoring function if for all  $x, y \in X$ ,

- (i)  $s(x, x) > 0$ ,
- (ii)  $s(x, x) \geq s(x, y)$ , and
- (iii)  $s(x, x) \geq s(y, x)$ .

Thus, a similarity map is sane if every element of  $\Sigma$  ‘keeps its identity’ with respect to it. Every point is similar to itself and this similarity cannot be smaller than similarity to any other point.

**Proposition 4.5.** *Let  $\Sigma$  be a set and let  $s : \Sigma \times \Sigma \rightarrow \mathbb{R}$  be a sane scoring function over  $\Sigma$ . Suppose  $\gamma, \delta \in \Gamma(\Sigma)$  and  $S$  the global similarity on  $\Sigma^*$  with respect to  $s, \delta$  and  $\gamma$ . Then,  $S$  is a sane scoring function and for all  $x \in \Sigma^*$ ,*

$$S(x, x) = \sum_{i=1}^{|x|} s(x_i, x_i). \quad (24)$$

□

Proposition 4.5 and Theorem 3.10 give us a straightforward way to convert global similarities to (quasi-) metrics. Since this transformation is based on the transformations of similarity scores to distances on generators, we first introduce additional nomenclature.

**Definition 4.6.** Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . For a sane scoring function  $s$  on  $\Sigma$ , we will use  $\text{AQ}^p(s)$  to denote the distance  $q$  on  $\Sigma$  given by

$$q(a, b) = (s(a, a) - s(a, b))^{1/p} \quad (25)$$

and  $\text{AM}^p(s)$  to denote the distance  $d$  on  $\Sigma$  given by

$$d(a, b) = (s(a, a) + s(b, b) - s(a, b) - s(b, a))^{1/p}. \quad (26)$$

Note that at this stage we do not make an assumption that  $\text{AQ}^p(s)$  is a quasi-metric nor that  $\text{AM}^p(s)$  is a metric.

**Corollary 4.7.** *Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . Suppose  $s$  is a sane scoring function on  $\Sigma$ ,  $d = \text{AQ}^p(s)$  is a quasi-metric on  $\Sigma$  and  $\gamma, \delta \in \Gamma_{CL}(\Sigma)$  such that*

$$\gamma(b) - \gamma(a) \leq d^p(a, b) \quad (27)$$

and

$$s(a, a) + \delta(a) - s(b, b) - \delta(b) \leq d^p(a, b). \quad (28)$$

*Let  $S$  be the global similarity with respect to  $s, \gamma$  and  $\delta$  and let  $\alpha(x) = \gamma(x)^{1/p}$  and  $\beta(x) = (S(x, x) + \delta(x))^{1/p}$  for all  $x \in \Sigma^+$ . Then, the  $\ell^p$  edit distance  $D = \text{EQ}^p(d, \alpha, \beta)$  is given for all  $x, y \in \Sigma^*$  by the formula*

$$D(x, y) = (S(x, x) - S(x, y))^{1/p}. \quad (29)$$

□

As with edit distances, we now introduce a nomenclature for quasi-metrics and metrics obtained from similarities.

**Definition 4.8.** Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . Suppose  $D$  is an  $\ell^p$  edit distance obtained from a global similarity  $S$  on  $\Sigma^*$  using the formula (29) of Corollary 4.7, where  $S$  extends  $s : \Sigma \times \Sigma$  and  $\gamma, \delta \in \Gamma_{\text{CL}}(\Sigma)$ . We will write  $D = \mathbf{GQ}^p(s, \gamma, \delta)$  if  $D$  is a quasi-metric and  $D = \mathbf{GM}^p(s, \gamma, \delta)$  if  $D$  is a metric.

The above nomenclature is redundant, in that every distance derived from similarities using Corollary 4.7 can be expressed using the nomenclatures for edit distances and distances on  $\Sigma$  introduced in Definition 4.6. We have chosen to nevertheless introduce the additional notation in order to emphasize that the distances on the free monoid are derived from similarities and also because the computation of distances can be performed using algorithms for similarities. This notation will also be convenient in the following sections, where local similarities are discussed.

**Example 4.9.** Let  $\Sigma$  be a set and suppose  $s$  is a sane symmetric function  $\Sigma \times \Sigma \rightarrow \mathbb{R}$  and  $\gamma \in \Gamma_{\text{CL}}(\Sigma)$ , depending only on length. This is a very frequent setup in pairwise comparison of DNA and protein sequences (see Section 6 below for more detailed discussion). Define for all  $a, b \in \Sigma$ ,  $s'(a, b) = 2s(a, b) - s(b, b)$  and for all  $x \in \Sigma^+$ ,  $\gamma'(x) = 2\gamma(x) + \sum_i s(x_i, x_i)$  and  $\delta'(x) = 2\gamma(x)$ .

Suppose that the distance  $d = \mathbf{AQ}^p(s') = \mathbf{AM}^p(s)$  is a metric on  $\Sigma$ . Since  $s$  is sane,  $s'$  is also sane and we have

$$|s'(a, a) - s'(b, b)| = |s(a, a) - s(b, b)| \leq d^p(a, b).$$

Therefore, since  $\gamma$  depends solely on length, the requirements (27) and (28) of Corollary 4.7 are satisfied. Let  $S$  be the global similarity extending  $s, \gamma$  and  $\gamma$  and let  $S'$  be the global similarity extending  $s', \gamma'$  and  $\delta'$ . We conclude that the distance  $D$  given by

$$D(x, y) = (S'(x, x) - S'(x, y))^{1/p} = (S(x, x) + S(y, y) - 2S(x, y))^{1/p} \quad (30)$$

is the metric  $\mathbf{GM}^p(s', \gamma', \delta')$ . This metric can also be expressed as  $\mathbf{EM}^p(\mathbf{AM}^p(s), \alpha)$ , where  $\alpha(x) = (S(x, x) + \gamma(x))^{1/p}$  for all  $x \in \Sigma^+$ .

## 5 Local Similarity

Local similarity is computed using the Smith-Waterman algorithm [69].

**Definition 5.1.** Let  $\Sigma$  be a set,  $s : \Sigma \times \Sigma \rightarrow \mathbb{R}$ , and let  $\gamma, \delta \in \Gamma(\Sigma)$ . Let  $x, y \in \Sigma^*$ ,  $m = |x|$  and  $n = |y|$ . The *Smith-Waterman* dynamic programming matrix, denoted  $\mathbf{SW}(x, y, s, \gamma, \delta)$ , is an  $(m+1) \times (n+1)$  matrix  $\mathbf{H}$  with rows and columns indexed from 0 such that  $\mathbf{H}_{0,0} = 0$  and for all  $1 \leq i \leq m$  and  $1 \leq j \leq n$ ,  $\mathbf{H}_{i,0} = 0$ ,  $\mathbf{H}_{0,j} = 0$  and

$$\mathbf{H}_{i,j} = \max \left\{ \mathbf{H}_{i-1,j-1} + s(x_i, y_j), \max_{1 \leq k \leq i} \{ \mathbf{H}_{i-k,j} - \delta(x_{i-k+1} \dots x_i) \}, \right. \\ \left. \max_{1 \leq k \leq j} \{ \mathbf{H}_{i,j-k} - \gamma(y_{j-k+1} \dots y_j) \}, \quad 0 \right\}.$$

The *local similarity* between the sequences  $x$  and  $y$  (given  $s$ ,  $\gamma$ , and  $\delta$ ), denoted  $H(x, y)$ , is defined to be the largest entry of  $\mathbf{H}$ , that is,  $H(x, y) = \max_{i,j} \mathbf{H}_{i,j}$ .

Local similarity between two words can be realized as global similarity of their fragments.

**Theorem 5.2** ([68]). *Let  $\Sigma$  be a set,  $s : \Sigma \times \Sigma \rightarrow \mathbb{R}$  and  $\gamma, \delta \in \Gamma(\Sigma)$ . Suppose  $S$  is a global similarity extending  $s, \gamma$  and  $\delta$  and  $H$  is the local similarity with respect to  $s, \gamma$  and  $\delta$ . Then, for all  $x, y \in \Sigma^*$ ,*

$$H(x, y) = \max_{\substack{x' \in \mathfrak{F}(x) \\ y' \in \mathfrak{F}(y)}} S(x', y'). \quad (31)$$

□

Although conversion of global similarities to distances outlined in Section 4 is relatively straightforward, its counterpart for local similarity is much less so. We now use the results from the previous sections to state our main result: construction of quasi-metrics which include conversions of local similarities.

**Theorem 5.3.** *Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . Let  $\rho$  be a separating quasi-metric on  $\Sigma^*$  that is arbitrarily decomposable of order  $p$ . Suppose  $f$  is a strictly positive and  $g$  is a non-negative function  $\Sigma \rightarrow \mathbb{R}$  and  $\bar{f}$  and  $\bar{g}$  are*

the canonical homomorphic extensions of  $f$  and  $g$ , respectively, to the free monoid  $\Sigma^*$ . Assume also that for all  $x, y \in \Sigma^*$ ,

$$\bar{f}(x) - \bar{f}(y) \leq \rho^p(x, y) \quad \text{and} \quad \bar{g}(y) - \bar{g}(x) \leq \rho^p(x, y). \quad (32)$$

Then, the function  $Q : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}$  defined by

$$Q(x, y) = \min_{\substack{\tilde{x} \in \mathfrak{F}(x) \\ \tilde{y} \in \mathfrak{F}(y)}} \left\{ \left( \bar{f}(x) - \bar{f}(\tilde{x}) + \bar{g}(y) - \bar{g}(\tilde{y}) + \rho^p(\tilde{x}, \tilde{y}) \right)^{1/p} \right\} \quad (33)$$

is a quasi-metric on  $\Sigma^*$ .

*Proof.* Let  $x, y, z \in \Sigma^*$ . Since  $\bar{f}(x) \geq \bar{f}(\tilde{x})$  and  $\bar{g}(y) \geq \bar{g}(\tilde{y})$  for any  $\tilde{x} \in \mathfrak{F}(x)$ ,  $\tilde{y} \in \mathfrak{F}(y)$  and since  $\rho$  is a quasi-metric and hence positive, it follows that  $Q(x, y) \geq 0$ . Furthermore, it is clear that  $Q(x, x) = 0$ .

Suppose that  $Q(x, y) = 0$ . Then, there exist  $\tilde{x} \in \mathfrak{F}(x)$  and  $\tilde{y} \in \mathfrak{F}(y)$  such that  $\bar{f}(x) - \bar{f}(\tilde{x}) + \bar{g}(y) - \bar{g}(\tilde{y}) + \rho^p(\tilde{x}, \tilde{y}) = 0$ . Since  $\rho(\tilde{x}, \tilde{y}) \geq 0$ ,  $\bar{f}(x) - \bar{f}(\tilde{x}) \geq 0$  and  $\bar{g}(y) - \bar{g}(\tilde{y}) \geq 0$  for any  $\tilde{x}, \tilde{y} \in \Sigma^*$ , it follows that  $\bar{f}(x) = \bar{f}(\tilde{x})$ ,  $\bar{g}(y) = \bar{g}(\tilde{y})$  and  $\rho(\tilde{x}, \tilde{y}) = 0$ . The first statement implies that  $x = \tilde{x}$  since  $f$  is a strictly positive function, while the last means that  $\tilde{x} = \tilde{y}$  (since  $\rho$  is a separating quasi-metric). Therefore,  $Q(x, y) = 0$  implies  $x \in \mathfrak{F}(y)$ . Hence,  $Q(x, y) = Q(y, x) = 0$  implies  $x \in \mathfrak{F}(y)$  and  $y \in \mathfrak{F}(x)$  and thus  $x = y$ .

To establish the triangle inequality suppose that

$$Q(x, y) = \left( \bar{f}(x) - \bar{f}(\tilde{x}) + \bar{g}(y) - \bar{g}(\tilde{y}) + \rho^p(\tilde{x}, \tilde{y}) \right)^{1/p} \quad (34)$$

for some  $\tilde{x} \in \mathfrak{F}(x)$ ,  $\tilde{y} \in \mathfrak{F}(y)$  and

$$Q(y, z) = \left( \bar{f}(y) - \bar{f}(\dot{y}) + \bar{g}(z) - \bar{g}(\dot{z}) + \rho^p(\dot{y}, \dot{z}) \right)^{1/p} \quad (35)$$

for some  $\dot{y} \in \mathfrak{F}(y)$  and  $\dot{z} \in \mathfrak{F}(z)$ . Write out  $\tilde{y} = y_i y_{i+1} \dots y_{i+m-1}$ ,  $\dot{y} = y_j y_{j+1} \dots y_{j+n-1}$  where  $m = |\tilde{y}|$ ,  $n = |\dot{y}|$ ,  $1 \leq i \leq i+m-1 \leq |y|$  and  $1 \leq j \leq j+n-1 \leq |y|$ . If  $\tilde{y}$  and  $\dot{y}$  overlap, that is, if  $i \leq j \leq m$  or  $j \leq i \leq n$ , let  $y'$  denote the whole overlapping fragment (for example, if  $i \leq j \leq i+m-1 \leq i+n-1$ ,  $y' = y_j y_{j+1} \dots y_{i+m-1}$  – see Figure 2). If  $\tilde{y}$  and  $\dot{y}$  do not overlap or either  $\tilde{y}$  or  $\dot{y}$  is identity, let  $y' = e$ .

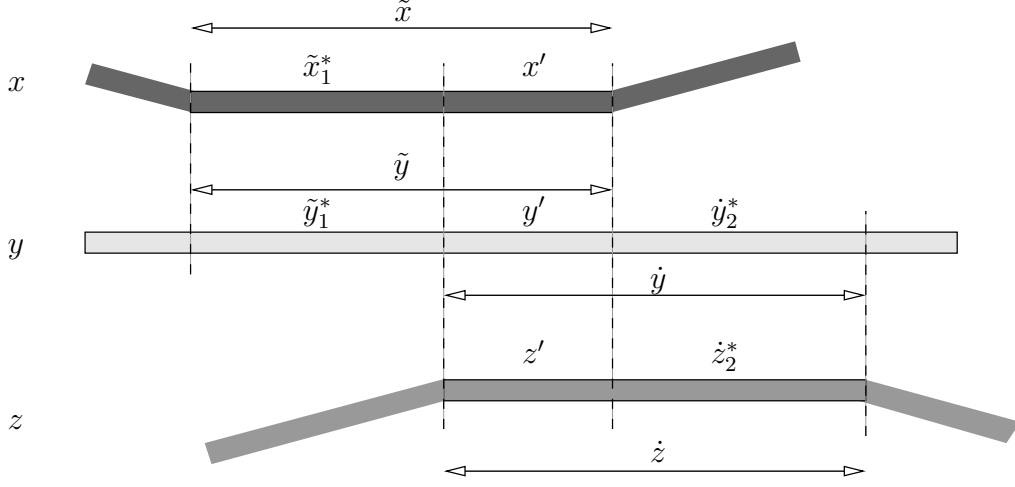


Figure 2: Decomposition of  $x$ ,  $y$  and  $z$ . In this pattern of overlap of  $\tilde{y}$  and  $\dot{y}$ , we have  $\tilde{x}_2^* = \tilde{y}_2^* = \dot{y}_1^* = \dot{z}_1^* = e$ .

Since  $\rho$  is arbitrarily decomposable of order  $p$ , there exist  $x', \tilde{x}_1^*, \tilde{x}_2^* \in \mathfrak{F}(\tilde{x})$  such that  $\tilde{x} = \tilde{x}_1^* x' \tilde{x}_2^*$  and  $\tilde{y}_1^*, \tilde{y}_2^*, u, v \in \mathfrak{F}(\tilde{y})$  such that  $\tilde{y} = \tilde{y}_1^* u y' v \tilde{y}_2^*$  and

$$\rho(\tilde{x}, \tilde{y}) \geq \left( \rho^p(\tilde{x}_1^*, \tilde{y}_1^*) + \rho^p(x', y') + \rho^p(\tilde{x}_2^*, \tilde{y}_2^*) \right)^{1/p}. \quad (36)$$

Furthermore, by the same assumption, there exist  $z', \dot{z}_1^*, \dot{z}_2^* \in \mathfrak{F}(\dot{z})$  such that  $\dot{z} = \dot{z}_1^* z' \dot{z}_2^*$  and  $\dot{y}_1^*, \dot{y}_2^*, \dot{u}, \dot{v} \in \mathfrak{F}(\dot{y})$  such that  $\dot{y} = \dot{y}_1^* \dot{u} y' \dot{v} \dot{y}_2^*$  and

$$\rho(\dot{y}, \dot{z}) \geq \left( \rho^p(\dot{y}_1^*, \dot{z}_1^*) + \rho^p(y', z') + \rho^p(\dot{y}_2^*, \dot{z}_2^*) \right)^{1/p}. \quad (37)$$

Therefore, using the Minkowski inequality,

$$\begin{aligned}
Q(x, y) + Q(y, z) &\geq \left( \bar{f}(x) - \bar{f}(\tilde{x}) + \bar{g}(y) - \bar{g}(\tilde{y}) \right. \\
&\quad \left. + \rho^p(\tilde{x}_1^*, \tilde{y}_1^*) + \rho^p(x', y') + \rho^p(\tilde{x}_2^*, \tilde{y}_2^*) \right)^{1/p} \\
&\quad + \left( \bar{f}(y) - \bar{f}(\dot{y}) + \bar{g}(z) - \bar{g}(\dot{z}) \right. \\
&\quad \left. + \rho^p(\dot{y}_1^*, \dot{z}_1^*) + \rho^p(y', z') + \rho^p(\dot{y}_2^*, \dot{z}_2^*) \right)^{1/p} \\
&\geq \left( \bar{f}(x) - \bar{f}(\tilde{x}) + \bar{f}(y) - \bar{f}(\dot{y}) + \rho^p(\tilde{x}_1^*, \tilde{y}_1^*) + \rho^p(\tilde{x}_2^*, \tilde{y}_2^*) \right. \\
&\quad \left. + \bar{g}(y) - \bar{g}(\tilde{y}) + \bar{g}(z) - \bar{g}(\dot{z}) + \rho^p(\dot{y}_1^*, \dot{z}_1^*) + \rho^p(\dot{y}_2^*, \dot{z}_2^*) \right. \\
&\quad \left. + (\rho(x', y') + \rho(y', z'))^p \right)^{1/p}.
\end{aligned}$$

Since  $\bar{f}$  and  $\bar{g}$  are additive functions that satisfy the inequality (32) and since  $y'$  is the full extent of the overlap between  $\tilde{y}$  and  $\dot{y}$ , we have

$$\begin{aligned}
&\bar{f}(x) - \bar{f}(\tilde{x}) + \bar{f}(y) - \bar{f}(\dot{y}) + \rho^p(\tilde{x}_1^*, \tilde{y}_1^*) + \rho^p(\tilde{x}_2^*, \tilde{y}_2^*) \\
&\geq \bar{f}(x) - \bar{f}(\tilde{x}) + \bar{f}(y) - \bar{f}(\dot{y}) + \bar{f}(\tilde{x}_1^*) - \bar{f}(\tilde{y}_1^*) + \bar{f}(\tilde{x}_2^*) - \bar{f}(\tilde{y}_2^*) \\
&\geq \bar{f}(x) - \bar{f}(x') \geq 0,
\end{aligned}$$

and

$$\begin{aligned}
&\bar{g}(y) - \bar{g}(\tilde{y}) + \bar{g}(z) - \bar{g}(\dot{z}) + \rho^p(\dot{y}_1^*, \dot{z}_1^*) + \rho^p(\dot{y}_2^*, \dot{z}_2^*) \\
&\geq \bar{g}(y) - \bar{g}(\tilde{y}) + \bar{g}(z) - \bar{g}(\dot{z}) - \bar{g}(\dot{y}_1^*) + \bar{g}(\dot{z}_1^*) - \bar{g}(\dot{y}_2^*) + \bar{g}(\dot{z}_2^*) \\
&\geq \bar{g}(z) - \bar{g}(z') \geq 0.
\end{aligned}$$

Hence, by the triangle inequality for  $\rho$ ,

$$Q(x, y) + Q(y, z) \geq \left( \bar{f}(x) - \bar{f}(x') + \bar{g}(z) - \bar{g}(z') + \rho^p(x', z') \right)^{1/p} \geq Q(x, z),$$

as required.  $\square$

**Remark 5.4.** We have shown in the separation part of the proof of Theorem 5.3 above that  $Q(x, y) = 0 \implies x \in \mathfrak{F}(y)$  and hence the associated partial order of the quasi-metric  $Q$  is  $x \leq_Q y \iff x \in \mathfrak{F}(y)$ . If  $g$  is a strictly positive function,  $Q$  is a separating quasi-metric and the partial order is

trivial:  $Q(x, y) = 0$  implies  $x = y$  and hence each point is only comparable to itself.

However, if  $g$  is zero everywhere, then  $Q(x, y) = 0$  and  $x \neq y$  implies that  $x$  is a factor of  $y$  while  $y$  is not a factor of  $x$ , so that  $x$  and  $y$  are non-trivially comparable. In this case, the quasi-metric  $Q$  is not separating and it generalizes the substring partial order: for every  $x, y \in \Sigma^*$  such that  $x \in \mathfrak{F}(y)$ , we have  $Q(x, y) = 0$ . Therefore,  $Q(x, y)$  can be interpreted as measuring how far is  $x$  from being a factor of  $y$ .

Since the identity  $e$  is a trivial factor of every word and  $\bar{f}(e) = 0$ , it follows that (in the case of  $g \equiv 0$ )  $Q(e, x) = 0$  for every  $x \in \Sigma^+$ , in contrast to  $\rho(e, x) \geq 0$ . On the other hand, it can be easily seen that  $Q(x, e) = (\bar{f}(x))^{1/p}$  and hence for all  $y \in \Sigma^+$ ,  $Q(x, y) \leq (\bar{f}(x))^{1/p} = Q(x, e)$ .

We now introduce a nomenclature for quasi-metrics and their associated metrics defined in Theorem 5.3.

**Definition 5.5.** Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . Suppose  $\rho$  is a separating quasi-metric on  $\Sigma^*$  and  $f$  and  $g$  are functions  $\Sigma \rightarrow \mathbb{R}$  that satisfy all the requirements of Theorem 5.3 with respect to  $\rho$  and  $p$ . Let  $Q$  be the quasi-metric obtained using the formula (33) of Theorem 5.3. We will write  $Q = \text{LQ}^p(\rho, f, g)$  if  $Q$  is a quasi-metric and  $Q = \text{LM}^p(\rho, f, g)$  if  $Q$  is a metric.

**Remark 5.6.** Edit distances described in Section 3 are always global: they measure the full cost of transformation between two words in  $\Sigma^*$ . Indeed, a truly ‘local’ distance, that is the distance measured on factors of words being compared, would not satisfy the triangle inequality.

The  $\text{LQ}^p$  distances are slightly different. The distance  $\rho$  contributes to  $Q$  by evaluating the pair of factors  $\tilde{x}$  and  $\tilde{y}$  that are ‘closest’ to each other (relative to  $\bar{f}$  and  $\bar{g}$ ), while  $\bar{f}$  and  $\bar{g}$  score the left-over pieces of  $x$  and  $y$ , respectively. The extent of  $\tilde{x}$  and  $\tilde{y}$  relative to  $x$  and  $y$  depends on the exact choice of functions  $f$  and  $g$  and their relation to the distance  $\rho$ . For example, when  $f$  and  $g$  are very large compared to  $\rho$ , the factors  $\tilde{x}$  and  $\tilde{y}$  will approach the whole sequences  $x$  and  $y$ . On the other hand, if  $f$  and  $g$  are small, they will contribute most to  $\text{LQ}^p(\rho, f, g)$ , depending on the exact properties of  $\rho$ .

When both  $f$  and  $g$  are strictly positive, the  $\text{LQ}^p$  distance has a global character in that the whole of  $x$  and  $y$  are accounted for. If  $g \equiv 0$ , only  $x$  contributes to the distance as a whole; the sequence  $y$  contributes only through its factor closest to a factor of  $x$ . In general, it is possible to favor  $x$  or  $y$  by appropriately choosing the values of  $f$  and  $g$ .



Theorem 5.3 can be applied to similarities in the following manner. Let  $Q = \text{LQ}^p(\rho, f, g)$ . Define a global similarity  $\sigma$  on  $\Sigma^*$  by

$$\sigma(x, y) = \bar{f}(x) + \bar{g}(y) - \rho^p(x, y). \quad (38)$$

Then,

$$\begin{aligned} Q(x, y) &= \left( \bar{f}(x) + \bar{g}(y) - \max_{\substack{\tilde{x} \in \mathfrak{F}(x) \\ \tilde{y} \in \mathfrak{F}(y)}} \{ \bar{f}(\tilde{x}) + \bar{g}(\tilde{y}) - \rho^p(\tilde{x}, \tilde{y}) \} \right)^{1/p} \\ &= \left( \bar{f}(x) + \bar{g}(y) - \max_{\substack{\tilde{x} \in \mathfrak{F}(x) \\ \tilde{y} \in \mathfrak{F}(y)}} \sigma(\tilde{x}, \tilde{y}) \right)^{1/p}. \end{aligned} \quad (39)$$

Hence, if  $\sigma$  can be computed using the Needleman-Wunsch algorithm (that is, if  $\rho$  is an  $\ell^p$  edit distance), then  $Q$  can always be evaluated by using the Smith-Waterman algorithm to compute the local similarity  $H(x, y) = \max\{\sigma(\tilde{x}, \tilde{y}) \mid \tilde{x} \in \mathfrak{F}(x), \tilde{y} \in \mathfrak{F}(y)\}$  and then using Equation (39).

Since the functions  $f$  and  $g$  as well as the quasi-metric  $\rho$  are arbitrary, the applicability of Theorem 5.3 to similarities is very wide. The following examples are simple corollaries of Theorem 5.3 and the results in Sections 3 and 4 that have important uses in computational biology.

**Example 5.7.** Let  $\Sigma$  be a finite set and suppose  $s$  is a sane symmetric function  $\Sigma \times \Sigma \rightarrow \mathbb{R}$  such that the distance  $d = \text{AQ}^1(s)$ , is a metric on  $\Sigma$ . Let  $\mu = \min\{s(a, b) \mid a, b \in \Sigma\}$  and let  $f(a) = s(a, a) - \mu$ . It is clear from the definitions of  $f$  and  $d$  that  $|f(a) - f(b)| \leq d(a, b) \leq f(a) + f(b)$ .

Let  $\rho$  be the arbitrarily decomposable metric extending the generalized Hamming distance based on  $d$  and  $f$  to  $\Sigma^*$ , as in Example 3.18 and define  $g : \Sigma \rightarrow \mathbb{R}$  by  $g(a) = s(a, a)$ . By Theorem 5.3 we can construct the distance  $\text{LQ}^1(\rho, g, g)$ , which is in fact the metric  $\text{LM}^1(\rho, g, g)$ . The underlying similarity  $\sigma$ , given by Equation (39), is

$$\sigma(x, y) = \begin{cases} 2s_n(x, y) & \text{if } |x| = |y| = n, \\ 2\mu & \text{if } |x| \neq |y|. \end{cases} \quad (40)$$

where  $s_n(x, y) = \sum_{i=1}^n s(x_i, y_i)$ . In computational biology applications,  $\mu$  will be negative (there will be at least two points in  $\Sigma$  that are dissimilar) and hence the local similarity will always be realized by aligning the fragments

of the same length. Therefore, the local similarity based on  $\sigma$  is *gapless similarity*, which has considerable historical importance since the first version of BLAST [1] suite of tools for sequence database search based on local similarities used a heuristic that computed gapless alignments. Gapless alignments had an advantage that they could be computed faster and the statistics of similarity scores arising from them were well characterized [38, 37].

In the following examples 5.8, 5.9 and 5.10, we will assume that  $s : \Sigma \times \Sigma \rightarrow \mathbb{R}$  is a sane scoring function,  $\gamma, \delta \in \Gamma_{\text{CL}}(\Sigma)$  only depend on length and  $S$  and  $H$  are global and local similarity with respect to  $s$ ,  $\gamma$  and  $\delta$ , respectively. In addition, let  $f(a) = s(a, a)$  for all  $a \in \Sigma$ .

**Example 5.8.** Suppose  $\text{AQ}^1(s)$  is a quasi-metric. By Corollary 4.7, the distance  $D$  on  $\Sigma^*$  given by  $D(x, y) = S(x, x) - S(x, y)$  is a quasi-metric  $\text{GQ}^1(s, \gamma, \delta)$ . Consider the distance  $Q = \text{LQ}^1(\text{GQ}^1(s, \gamma, \delta), f, 0)$ . It is easy to see that  $\bar{f}(x) = S(x, x) = H(x, x)$  and hence

$$Q(x, y) = S(x, x) - \max_{\substack{\tilde{x} \in \mathfrak{F}(x) \\ \tilde{y} \in \mathfrak{F}(y)}} S(\tilde{x}, \tilde{y}) = H(x, x) - H(x, y). \quad (41)$$

As remarked earlier, the partial order associated with  $Q$  in this case is subfragment partial order. Furthermore, the triangle inequality for  $Q$  is equivalent to

$$H(x, y) + H(y, z) \leq H(y, y) + H(x, z). \quad (42)$$

If  $H$  is symmetric, that is, if  $s$  is symmetric and  $\gamma = \delta$ , we have

$$Q(x, y) + H(y, y) = Q(y, x) + H(x, x), \quad (43)$$

and hence  $Q$  is a co-weightable quasi-metric and  $-H$  is a partial metric. Note that in this case, the triangle inequality (42) is exactly equivalent to the triangle inequality for the symmetrization  $M(x, y) = Q(x, y) + Q(y, x)$  (Example 5.10), that is, if  $M$  is a metric then  $Q$  is a quasi-metric.

The fact that Equation (41) gives a quasi-metric was first established in [73]. Indeed, the two generate equivalent neighborhoods: for any  $x \in \Sigma^*$ , the set of all points  $y \in \Sigma^*$  such that  $H(x, y) > \kappa$  is equal to the set  $\{y \in \Sigma^* : Q(x, y) < \varepsilon\}$  where  $\varepsilon = S(x, x) - \kappa$ .

**Example 5.9.** Recall the notation from Example 4.9, where  $s$  is symmetric,  $\gamma = \delta$ ,  $s'(a, b) = 2s(a, b) - s(b, b)$ ,  $\gamma'(x) = 2\gamma(x) + \sum_i s(x_i, x_i)$  and  $\delta'(x) =$

$2\gamma(x)$ . Let  $S'$  and  $H'$  be global and local similarity with respect to  $s'$ ,  $\gamma'$  and  $\delta'$ , respectively.

Suppose that  $\text{AQ}^p(s')$  is a quasi-metric (equivalently that  $\text{AM}^p(s)$  is a metric) and consider the quasi-metric  $Q' = \text{LQ}^p(\text{GM}^p(s', \gamma', \delta'), f, 0)$ . By the argument of Example 5.8,

$$Q'(x, y) = (H'(x, x) - H'(x, y))^{1/p} = (S(x, x) - H'(x, y))^{1/p}. \quad (44)$$

However, in this case the local similarity

$$H'(x, y) = \max_{\tilde{x}, \tilde{y}} S'(\tilde{x}, \tilde{y}) = \max_{\tilde{x}, \tilde{y}} (2S(x, y) - S(y, y)) \quad (45)$$

is clearly asymmetric. This similarity score has, to our knowledge, never been previously used for sequence comparison, although it can be easily computed using Smith-Waterman algorithm (provided that the particular implementation used allows composition-length gap penalties). It has the advantage that it is still true that  $H'$  is topologically equivalent to  $Q'$  and that  $Q'$  corresponds to the subfragment partial order.

The asymmetry of  $H'$  may be exploited to favor the integrity of one sequence over the other in biological sequence alignments. For example, in cases where translated DNA sequences are compared to proteins, it is desirable to emphasize the protein sequence, which is ‘real’ (experimentally established), at the expense of translated DNA sequences, which is only hypothetical. We intend to evaluate the broad utility of using variants of  $H'$  and  $Q'$  for biological sequence comparisons in a subsequent publication.

**Example 5.10.** Making the same assumptions as in Example 5.9 above, consider the metric  $M = \text{LM}^p(\text{GM}^p(s', \gamma', \delta'), f, f)$ . It is easy to see that  $M$  is indeed a metric given by

$$M(x, y) = (H(x, x) + H(y, y) - 2H(x, y))^{1/p}. \quad (46)$$

Equation (46), for  $p = 1$ , was extensively considered in computer science and computational biology. The LCS similarities (Examples 3.15 and 4.2) are related to distances in this way. Linial *et al.* [47] proposed using  $M$  as a distance on sets of protein sequences but did not explicitly prove it was a metric. Spiro and Macura [71] have given the conditions under which  $M$  is indeed a metric. Since  $H$  is here assumed symmetric, this result is equivalent to  $\text{LQ}^1(\text{GQ}^1(s, \gamma, \delta), f, 0)$  being a quasi-metric (Example 5.8), established by

Stojmirović [73] under slightly different assumptions. Itoh *et al.* [33] derived the same result as a corollary of a more general inequality for similarities that relied on the finiteness of the generator alphabet. In a poster abstract [16], Fischer proposed the general form of Equation (46) with arbitrary  $p$  as a way to convert similarities to distances and stated without proof the conditions for  $M$  to be a metric.

For  $p = 2$ , the form of Equation (46) resembles the formula for the canonical metric in inner-product vector spaces. In this case,  $x$  and  $y$  would be vectors and  $H$  would be a positive-definite bilinear form.

Theorem 5.3 can be applied in the context of free abelian monoids with no change. We illustrate this by a very simple, followed by a more biologically relevant example.

**Example 5.11.** Let  $\Sigma$  be the set of all prime numbers and let  $\Sigma^*$  be the free abelian monoid over  $\Sigma$  under multiplication (i.e. the set of natural numbers  $\mathbb{N}$ ). Let  $d_{\dagger}$  be a discrete metric on  $\Sigma$  (here we implicitly assume that  $\Sigma$  includes 1) and let  $f(a) = 1$  and  $\alpha(a) = 0$  for all  $a \in \Sigma$ . Let  $\rho$  be the Sellers-Graev metric extension of  $d_{\dagger}$  to  $\mathbb{N}$ . It is clear that  $\rho(x, y)$  is just the number of different prime factors between  $x$  and  $y$  (the non-matching prime factors are matched to 1) and that it is arbitrarily decomposable. Hence, we can apply Theorem 5.3 to obtain a quasi-metric  $Q$ , so that  $Q(x, y)$  is the number of prime factors of  $x$  not in common to  $y$ . The global similarity  $\sigma$  on  $\mathbb{N}$  (here equivalent to local similarity), given by  $\sigma(x, y) = \bar{f}(x) - \rho(x, y)$  evaluates to the number of common prime factors (excluding 1) between  $x$  and  $y$ .

**Example 5.12.** Let  $\Sigma$  be a finite set and let  $A(\Sigma^k)$  denote the free abelian monoid generated by the set of all words of length exactly  $k$  (we will call  $z \in \Sigma^k$  a  $k$ -tuple). Members of  $A(\Sigma^k)$  are therefore multisets of  $k$ -tuples. Now consider the same structure as in the previous example.

Let  $d_{\dagger}$  be a discrete metric on  $\Sigma^k \cup \{e\}$  and let  $f(a) = 1$  and  $\alpha(a) = 0$  for all  $a \in \Sigma$ . Let  $\rho$  be the Sellers-Graev metric extension of  $d_{\dagger}$  to  $A(\Sigma^k)$ . All requirements of Theorem 5.3 still apply. The value  $Q(x, y)$  is the number of  $k$ -tuples that are contained in  $x$  but not in  $y$  and the global (and local) similarity  $\sigma$  gives the number of  $k$ -tuples common to both  $x$  and  $y$ .

The similarity  $\sigma$  has been used in computational biology as a computationally inexpensive approximation of global similarity between two sequences [39, 12]. Each sequence is mapped to  $A(\Sigma^k)$  by taking the multiset of all of its (overlapping)  $k$ -tuples and the similarity  $\sigma$  is used to approximate the global similarity  $S$ .

## 6 Scoring Functions on Generators

In the previous sections we have made no assumption on the set of generators  $\Sigma$  and all our results apply to arbitrary sets. However, as we noted before, the principal objects motivating our results are sets of biological sequences and profiles derived from them. The former two sets are finite and therefore the scoring functions over them are given by *score matrices*. We therefore proceed to discuss the similarity and distance measures on the sets of nucleotides, amino acids and profiles and their applicability to our theory.

### 6.1 Nucleotide scoring matrices

The nucleotide alphabet consists of only 4 letters (A, C, G, and T) and the score matrices most frequently used for database search depend on only two parameters, for scoring a match or a mismatch of two nucleotides. For example, the *blastn* program, a part of the BLAST [2] suite of tools for sequence database search based on local similarities, which searches a DNA database with a DNA sequence as a query, uses the scoring matrix of the form

$$s(a, b) = \begin{cases} 5 & \text{if } a = b \\ -4 & \text{if } a \neq b. \end{cases} \quad (47)$$

The above scoring function is obviously sane and the distance  $d = \mathbf{AQ}^p(s)$  is a discrete metric for any  $1 \leq p < \infty$ . Therefore, all match/mismatch scoring schemes satisfy the requirements of Theorem 3.10 and its corollaries.

More complex score matrices, where transitions (changes  $\mathbf{C} \leftrightarrow \mathbf{T}$  and  $\mathbf{A} \leftrightarrow \mathbf{G}$ ) have different scores than transversions (all other mutations) have been proposed for improving the accuracy of database searches [72, 10]. It is easy to show that the distance  $\mathbf{AQ}^1(s)$  (and hence  $\mathbf{AQ}^p(s)$  for all  $p$ ) will still satisfy the triangle inequality and hence be a metric if the value of distance associated by transition is not greater than twice the transversion distance. Since the likelihood and hence the similarity score of transition is larger than that of transversion, this condition is very likely to be satisfied in practice. For example, all scoring matrices examined by States *et al.* [72] satisfy this condition and are sane.

## 6.2 Amino acid scoring matrices

Unlike the nucleotide alphabet, the standard amino acid alphabet consists of 20 amino acids of markedly different chemical properties and structural roles. Hence, the regularly used amino acid scoring matrices are much more complex than the matrices over nucleotides discussed above. Many amino acid scoring matrices were developed over the years for various purposes, including sequence similarity search, structural prediction and phylogenetic analysis [54, 77, 40]. Most of them arise from analysis of sets of peptide sequences known to be to a certain extent related.

Dayhoff *et al.* [9] proposed in 1970s the family of scoring matrices called PAM, which were based on a Markov model of evolution of proteins. PAM matrices were the original standard choice for sequence comparison. Several improved versions of PAM matrices were constructed later [20, 34, 52, 53, 78], in order to address some of the deficiencies arising from lack of sufficient data at the time of the construction of the original PAM family. For PAM-like matrices, the larger the number appended to their name (such as PAM- $n$ ), the sequences to be compared are assumed to have more diverged in evolution.

Presently, the most widely used family of scoring matrices is BLOSUM, derived by Henikoff and Henikoff in 1992 [28] using an empirical procedure. In particular, the BLOSUM62 matrix has long been believed to be among the best performing matrices for general sequence similarity search [29] and is used as default by BLAST (more specifically, the *blastp* program). In contrast to the PAM-like matrices, the larger the number appended to the name of a BLOSUM matrix, the more the sequences to be compared are assumed to be closely related.

In addition to the above mentioned families, some score matrices were constructed specifically for searches involving transmembrane regions of proteins [35, 51, 56] while others were derived from structural alignments in order to improve sensitivity of searches involving distantly related proteins [63, 36, 5].

Table 1 shows the numbers of violations of the triangle inequality for the distances  $AQ^1$ ,  $AQ^2$  and  $AM^2$  obtained from several common (symmetric) score matrices. The matrices featured in Table 1 are all sane and represent only a very small sample of all existing amino acid score matrices that are most frequently used and cited.

All of the scoring matrices mentioned so far were symmetric with the exception of the SLIM family [51] for comparison of transmembrane proteins.

| Matrix      | Reference | AQ <sup>1</sup> | AQ <sup>2</sup> | AM <sup>2</sup> |
|-------------|-----------|-----------------|-----------------|-----------------|
| PAM40       | [9]       | 28              | 0               | 0               |
| PAM120      | [9]       | 88              | 0               | 0               |
| PAM250      | [9]       | 168             | 21              | 0               |
| GONNET      | [20]      | 144             | 0               | 0               |
| BLOSUM45    | [28]      | 0               | 0               | 0               |
| BLOSUM50    | [28]      | 0               | 0               | 0               |
| BLOSUM62    | [28]      | 0               | 0               | 0               |
| BLOSUM80    | [28]      | 0               | 0               | 0               |
| JTT         | [34]      | 170             | 34              | 34              |
| JTTtm       | [35]      | 214             | 18              | 20              |
| BC0030      | [5]       | 214             | 12              | 4               |
| SDM         | [63]      | 134             | 0               | 0               |
| HSDM        | [63]      | 142             | 6               | 0               |
| OPTIMA      | [36]      | 74              | 15              | 2               |
| PHAT75/73   | [56]      | 6               | 0               | 0               |
| VTML160     | [52]      | 28              | 0               | 0               |
| VTML250     | [52]      | 100             | 14              | 0               |
| dist.20comp | [7]       | 0               | 0               | 0               |
| PMB120      | [78]      | 0               | 0               | 0               |
| PMB250      | [78]      | 8               | 3               | 0               |

Table 1: Number of triples of amino acids failing the triangle inequality for distances derived from various symmetric score matrices. All the matrices are considered over the standard (20 letter) amino acid alphabet (that is, excluding non-standard letters representing more than one amino acid). Due to symmetry of similarity scores, the triangle inequalities for AQ<sup>1</sup> and AM<sup>1</sup> are equivalent and the column for AM<sup>1</sup> is omitted.

Yu *et al.* [87] recently proposed a concept of *compositionally adjusted score matrices*, which are asymmetric and which can be derived from symmetric score matrices by considering different background frequencies of amino acids in the first vs. the second sequence. The rationale for compositional adjustment is that some proteins, especially from organisms with biased amino acid usage, can have significantly different background frequencies of amino acids, than the ones used to construct the standard matrices. It was demonstrated in [87] that using compositional adjustment results in improvement of sensitivity of pairwise sequence comparison.

| Matrix      | <i>C. tetani</i> |                 |                 |                 | <i>M. tuberculosis</i> |                 |                 |                 |
|-------------|------------------|-----------------|-----------------|-----------------|------------------------|-----------------|-----------------|-----------------|
|             | AQ <sup>1</sup>  | AQ <sup>2</sup> | AM <sup>1</sup> | AM <sup>2</sup> | AQ <sup>1</sup>        | AQ <sup>2</sup> | AM <sup>1</sup> | AM <sup>2</sup> |
| PAM40       | 36               | 0               | 36              | 0               | 40                     | 0               | 40              | 0               |
| PAM120      | 129              | 0               | 126             | 0               | 113                    | 0               | 116             | 0               |
| GONNET      | 152              | 0               | 152             | 0               | 151                    | 0               | 150             | 0               |
| BLOSUM45    | 0                | 0               | 0               | 0               | 4                      | 0               | 4               | 0               |
| BLOSUM50    | 1                | 0               | 2               | 0               | 3                      | 0               | 2               | 0               |
| BLOSUM62    | 1                | 0               | 2               | 0               | 1                      | 0               | 2               | 0               |
| BLOSUM80    | 0                | 0               | 0               | 0               | 0                      | 0               | 0               | 0               |
| JTT         | 353              | 11              | 378             | 0               | 320                    | 5               | 330             | 0               |
| BC0030      | 234              | 3               | 244             | 4               | 249                    | 2               | 272             | 4               |
| SDM         | 132              | 0               | 132             | 0               | 132                    | 8               | 132             | 0               |
| HSDM        | 144              | 1               | 144             | 0               | 143                    | 0               | 142             | 0               |
| OPTIMA      | 77               | 4               | 78              | 2               | 78                     | 2               | 80              | 2               |
| PHAT75/73   | 10               | 0               | 12              | 0               | 19                     | 0               | 26              | 0               |
| VTML160     | 32               | 0               | 34              | 0               | 42                     | 0               | 50              | 0               |
| dist.20comp | 0                | 0               | 0               | 0               | 0                      | 0               | 0               | 0               |
| PMB120      | 0                | 0               | 0               | 0               | 0                      | 0               | 0               | 0               |

Table 2: Number of triples of amino acids failing the triangle inequality for various compositionally adjusted asymmetric score matrices. Each matrix was adjusted from a symmetric matrix by using the composition of either *C. tetani* or *M. tuberculosis* proteome as the first set of frequencies, together with the implicit amino acid frequencies from BLOSUM62 as the second set of frequencies.

Table 2 shows the violations of the triangle inequality for the distances obtained from some of the matrices from Table 1, adjusted to take into account the amino acid compositions of proteomes of bacterial species *Clostridium tetani* and *Mycobacterium tuberculosis*. Both of these species have compositionally biased genomes and proteomes. The matrices were constructed using a Newtonian procedure described in [86] and [3]. The background distribution for the second sequence comes from the original BLOSUM62 matrix. In this way, the constructed similarity scores and distances can be used to compare sequences known to come from the above organisms to sequences from general datasets.

Table 1 and Table 2 demonstrate that most scoring matrices, both symmetric and asymmetric, can be converted to the AM<sup>2</sup> metric while many can



be converted to  $AQ^2$  quasi-metric as well. In contrast, most matrices fail the triangle inequalities for  $AQ^1$  and  $AM^1$ . Therefore, our generalization of edit distances and related sequence similarities to  $\ell^p$  form allows us to use a much wider class of matrices to construct (quasi-) metrics on the set of all protein sequences. This is in contrast to the  $\ell^1$ -type results from the previous work [73, 71], which only apply to the BLOSUM family plus a few more similar matrices.

### 6.3 Profiles

Recall (Example 2.2) that given a set  $\Sigma$ , a profile over  $\Sigma$  is a word in the free monoid  $\mathcal{M}(\Sigma)^*$ , that is, a finite sequence of finite measures over  $\Sigma$ . In biological applications,  $\Sigma$  is finite and therefore a profile  $x$  can be treated as a sequence of vectors  $\mathbf{x}_i \in \mathbb{R}^n$ , where  $n = |\Sigma|$ . For each  $i$ , the vector  $\mathbf{y} = \mathbf{x}_i$  has non-negative entries. In some applications, it is further assumed that  $\mathbf{y}$  is a probability distribution, that is, that  $\sum_j y_j = 1$ .

In biological context, profiles represent generalized sequences over the basic alphabet  $\Sigma$  where each position has a probability distribution of letters instead of a single letter. They were originally introduced by Gribskov *et al.* [24] in order to improve sensitivity of homology search by considering the information contained in multiple alignments of related proteins to query sequence databases. To do so, a *Position Specific Score Matrix* or PSSM, which gives a similarity score for each letter in  $\Sigma$  for each position in the query profile, is constructed. The profile-sequence comparison using PSSM can then be performed using the dynamic programming algorithms such as Needleman-Wunsch or Smith-Waterman. Profiles can also be used directly in probabilistic Hidden Markov Models [11]. Profile-based homology searches are widely used and have been shown in general to be more sensitive than sequence database searches with normal sequences as queries [2, 11].

Profiles can also be compared to other profiles as members of the free monoid  $\mathcal{M}(\Sigma)^*$  using distances or similarities discussed in Sections 3, 4 and 5: all that is necessary is to assign a distance or similarity measure on  $\mathcal{M}(\Sigma)$  and gap penalties. Many scoring schemes were proposed in due course and we present only a few examples below. For a more detailed overview we refer the reader to the papers of Edgar and Sjölander [13] and Marti-Renom *et al.* [49], which study their performance for aligning distantly related protein sequences.

Let  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  be two measures in  $\mathcal{M}(\Sigma)$  and let  $\hat{s}$  and  $\hat{d}$  denote a similarity

and a distance function, respectively. The symbol  $\|\cdot\|$  denotes the  $\ell^2$  norm on  $\mathbb{R}^n$ .

**Example 6.1.** The simplest similarity score between two vectors, used in CLUSTALW software for multiple sequence alignment [76] (see also Section 7) is to compute their average over a score matrix  $s$  on  $\Sigma$ :

$$\hat{s}(\mathbf{x}, \mathbf{y}) = \sum_i \sum_j x_i y_j s(x_i, y_j). \quad (48)$$

In general,  $\text{AQ}^p(\hat{s})$  and  $\text{AM}^p(\hat{s})$  are not a quasi-metric or a metric, respectively.

**Example 6.2.** A natural candidate for similarity score between two vectors  $\mathbf{x}$  and  $\mathbf{y}$  is their dot product, used in [65]:

$$\hat{s}(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} = \sum_j x_j y_j. \quad (49)$$

Clearly,  $\hat{d} = \text{AM}^2(\hat{s})$  is the standard Euclidean distance:

$$\hat{d}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_j (x_j - y_j)^2}. \quad (50)$$

**Example 6.3.** A variation of the above is the correlation coefficient or cosine of the angle between two vectors used in the LAMA algorithm [62]:

$$\hat{s}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum_j x_j y_j}{\sqrt{\sum_j x_j^2 \sum_j y_j^2}}. \quad (51)$$

Here  $\hat{d} = \text{AM}^2(\hat{s})$  can be easily shown to satisfy the triangle inequality. In general,  $\hat{d}$  does not separate points, but if  $\mathbf{x}$  and  $\mathbf{y}$  are assumed to be probability vectors, then  $\hat{d}$  is indeed a metric.

**Example 6.4.** The Jensen-Shannon divergence between two probability vectors  $\mathbf{x}$  and  $\mathbf{y}$ , denoted  $D^{\text{JS}}$  is given by

$$D^{\text{JS}}(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_i \left[ x_i \log \frac{2x_i}{x_i + y_i} + y_i \log \frac{2y_i}{x_i + y_i} \right]. \quad (52)$$

While  $D^{\text{JS}}$  is not a metric, taking the square root, that is, letting  $\hat{d}(\mathbf{x}, \mathbf{y}) = \sqrt{D^{\text{JS}}(\mathbf{x}, \mathbf{y})}$  does give a metric [14]. Yu [85] proposed using this metric to compare probability distributions that are components of profiles while Yona and Levitt [84] used the following similarity score:

$$\hat{s}(\mathbf{x}, \mathbf{y}) = \left(1 - D^{\text{JS}}(\mathbf{x}, \mathbf{y})\right) \left(1 + D^{\text{JS}}\left(\frac{\mathbf{x} + \mathbf{y}}{2}, \boldsymbol{\pi}\right)\right), \quad (53)$$

where  $\boldsymbol{\pi}$  denotes a background distribution.

The above examples suggest that  $\ell^2$ -type edit distances and global and local similarities arising from them, could be appropriate for profile-profile comparisons.

## 7 Applications and Future Directions

Our results provide a way to construct a large variety of metrics and quasi-metrics on free semigroups. In particular, we are able to extend the conversion of similarity score matrices into alphabet (generator) distances, to the corresponding conversions of sequence similarities, global and local, to sequence distances. Hence, we are able to treat biosequence sets as spaces with geometry. The metric and quasi-metric structures provide a much richer framework than the topologies induced from them: for biosequences,  $\Sigma$  is finite and hence all topologies induced from  $\ell^p$  edit distances or local similarity (quasi-) metrics are equivalent to the discrete topology.

In terms of statistical characterization, since we allowed more general gap penalties and asymmetric scoring matrices, the established statistics for similarities may not be fully transferred to our general distances. For this reason, to fully exploit our general formulation, it is important to further elaborate on its statistical aspects, which is beyond the scope of the current paper.

Apart from setting a general geometric framework for sequence comparison, most direct applications to biology involve clustering. For example, global clustering of protein sequences has been performed [47, 66], using the metric from Example 5.10 and other derivations from similarity score. However, these works did not consider quasi-metrics and partial orders that could provide a more accurate view of the global protein sequence space. Applications to indexing and multiple sequence alignment, which we discuss in more detail below, can also be considered as clustering.

## Indexing for database search

One of the principal motivations for establishing the triangle inequalities for similarity scores in the literature [73, 71, 33] was to accelerate similarity search of large DNA and protein sequence databases. It has been identified early on that using the full Needleman-Wunsch and Smith-Waterman dynamic programming algorithms to search sequence datasets by sequentially scanning all entries is prohibitively computationally expensive and heuristic methods such as FASTA [57] and BLAST [1, 2] were developed. While very fast, these methods are not *consistent* [61], that is, they are not guaranteed to retrieve all true neighbors of a given query point. Furthermore, both FASTA and BLAST sequentially scan all of the sequences in the dataset being searched. The idea behind using the triangle inequalities for accelerating similarity search is to use the intrinsic ‘geometry’ of the dataset and the space it lies in to construct an *indexing scheme* [27, 61], a structure that allows fully retrieving a similarity query without scanning the whole dataset. A large amount of effort was spent on producing efficient indexing structures, principally concentrating on datasets that are equipped with a metric or a vector space structure: a good overview is by Hjalton and Samet in [30].

Let  $X \subset \Sigma^*$  be a finite sequence dataset. A range query of  $X$  based on local similarity  $H$  (depending on the score matrix  $s$  and gap penalties  $\gamma$  and  $\delta$ ), centered at the query point  $x \in \Sigma^*$  with threshold  $\kappa$  is the set

$$\mathcal{Q}_H(x, \kappa) = \{y \in X : H(x, y) \geq \kappa\}. \quad (54)$$

We will now consider some ways to construct an indexing structures that accelerate retrieval of  $\mathcal{Q}_H$ .

The first way is to consider biological sequences purely as strings with simple similarity measures often related to Levenstein distance and use string-based techniques such as hashing [19, 6, 41, 75] or suffix arrays [32, 31]. Such indexing schemes are often not consistent but may show good performance on datasets of DNA sequences where the similarity measure is very simple. For proteins, one approach was to construct a biologically meaningful metric on the amino acid alphabet and use the edit distance extension of it for sequence comparison and indexing [48]. This has an advantage that existing methods for indexing metric spaces can be directly applied but ignores the need for local similarities, which cannot be converted into edit distances.

The other approach, investigated by Spiro and Macura [71] and more thoroughly implemented by Itoh *et al.* [33], was to use the inequality (42),

which holds for some amino acid scoring matrices (Table 1). The idea is to cluster proteins according to the local similarity score  $H$  or associated metric  $\text{LM}^1(\text{GM}^1(s', \gamma', \delta'), f, f)$  (where similarity score is assumed symmetric and  $f(a) = s(a, a)$  – see Example 5.10), and then, when searching, to compare the query sequence to centers of clusters first and only scan those clusters that overlap the query.

Note that while the neighborhoods of  $\text{LQ}^1(\text{GQ}^1(s, \gamma, \delta), f, 0)$  are indeed equivalent to queries  $\mathcal{Q}_H$ , this is no longer true for neighborhoods of its metric symmetrization  $\text{LM}^1(\text{GM}^1(s', \gamma', \delta'), f, f)$ . Hence, direct indexing with respect to the local similarity metric may not be optimal. Furthermore, not all similarity score matrices give rise to  $\ell^1$  quasi-metrics  $\text{AQ}^1(s)$  (Table 1). Many more can be converted to  $\text{AM}^2(s)$  and hence give rise to metrics  $\text{LM}^2(\text{GM}^2(s', \gamma', \delta'), f, f)$ . Profile-profile comparison methods, relying on inner product for the distance between two distributions, also naturally induce  $\ell^2$ -type distances. None of the methods described above can efficiently cope with this situation and yet there exists a simple way to convert such similarity queries to a sequence of metric queries.

Suppose  $M(x, y) = (H(x, x) + H(y, y) - 2H(x, y))^{1/p}$  is a metric for some symmetric local similarity  $H$ . Let

$$Z_\xi = \{x \in \Sigma^* : H(x, x) = \xi\}. \quad (55)$$

We call each set  $Z_\xi$  a *fiber* and it is obvious that  $\Sigma^*$  is a disjoint union of all  $Z_\xi$ , where  $\xi$  runs over the range of self-similarities. For our applications, this range is finite because the sequence datasets are finite. Now consider a query  $\mathcal{Q}_H(x, \kappa)$  and let  $\varepsilon(x, \xi, \kappa) = (H(x, x) + \xi - 2\kappa)^{1/p}$ . It is easily established that

$$\mathcal{Q}_H(x, \kappa) = \bigsqcup_{\xi} \overline{\mathfrak{B}}(x, \varepsilon(x, \xi, \kappa))|_{Z_\xi}, \quad (56)$$

where  $\overline{\mathfrak{B}}(x, \varepsilon(x, \xi, \kappa)) = \{y \in X : M(x, y) \leq \varepsilon(x, \xi, \kappa)\}$  (the closed ball of radius  $\varepsilon(x, \xi, \kappa)$  about  $x$ ).

Hence, to process each local similarity range query, it is sufficient to process a metric range query  $\overline{\mathfrak{B}}(x, \varepsilon(x, \xi, \kappa))$  on each fiber and then collect the results. For practical purposes the fibers need to be reasonably large and small in number, but that is often true because the score matrices are integer-valued. Adjacent fibers that contain too few points can be merged if care is exercised when collecting final results. Each fiber can be indexed separately as a metric space with one of the many existing access methods

[30] or by using a new technique. The decomposition (56) was proposed in  $\ell^1$  form in [74] for indexing similarity-based range queries and was in turn inspired by decomposition of weightable quasi-metric spaces into fibers used by Vitolo [79].

Therefore, using fibers, a consistent indexing scheme can be constructed for most existing local similarity measures on biological sequences and profiles. The performance of such schemes is not guaranteed – it depends on the exact geometry of sequence datasets [60, 61]. Hence, our theoretical results represent only the first step towards efficient and consistent access methods that are to be achieved in future.

An alternative to fiber decomposition for cases where  $\mathbf{LQ}^p(\mathbf{GQ}^p(s, \gamma, \delta), f, 0)$  is truly a quasi-metric is to use the quasi-metric directly to index the dataset. Pestov and Stojmirović [61] proposed the concept of a quasi-metric tree: a general indexing scheme for retrieving queries based on quasi-metrics and established conditions for its consistency. Note that in the  $\ell^1$  case, using inequality (42) directly, as in [33], produces a structure that is equivalent to a quasi-metric tree.

## Progressive multiple sequence alignment

Multiple sequence alignment (MSA) is among the most valuable tools in computational biology. It allows extracting and representing biologically important commonalities from sets of sequences [25]. Construction of multiple alignments from sets of sequences has been extensively researched and a variety of techniques have been proposed [25, 10]. The full dynamic programming algorithm for MSA is NP-complete [80] and therefore heuristics are commonly employed. One popular heuristic approach is progressive alignment [15]. First, a guide tree is constructed from pairwise dissimilarities between sequences. Then, larger and larger groups of sequences are aligned in pairwise manner, following the branching order of the guide tree from the leaves towards the root. A number of popular software packages for MSA of protein sequences [76, 39, 12, 45] implement this heuristics.

The success of this approach, greedy in nature, crucially depends on a faithful and evolutionarily meaningful construction of a guide tree for the set of sequences to be aligned. When constructing their guide trees, most methods do not use a true metric to compute pairwise distances [76, 39, 12], while those that do [45], use the Levenshtein distance, overlooking the similarities between closely related amino acids.

There are advantages in using the true metric distance for agglomerative hierarchical clustering. For example, the triangle inequality ensures the transitivity of closeness in distance measure. Furthermore, when this is the case, it was shown that the difference between a hierarchical clustering and the optimal  $k$ -clustering is bounded [8].

In this paper we have demonstrated a way to construct a large class of (quasi-)metric distances from similarity scores that also naturally account for functional relatedness among amino acids. The quasi-metrics developed in Section 5 can also provide a rigorous way to naturally interpolate from global to local similarities in constructing guide trees.

### Embeddings into vector spaces

Let  $Q^u$  be the metric symmetrizing the quasi-metric  $Q = \text{LQ}^p(\rho, f, 0)$ , where  $Q^u(x, y) = Q(x, y) + Q(y, x)$  for all  $x, y \in \Sigma^*$ . Observe that by the triangle inequality for  $Q$ ,

$$(\bar{f}(x))^{1/p} - (\bar{f}(y))^{1/p} = Q(x, e) - Q(y, e) \leq Q(x, y), \quad (57)$$

and hence, letting  $\alpha(x) = (\bar{f}(x))^{1/p}$ , we have

$$|\alpha(x) - \alpha(y)| \leq Q^u(x, y) \leq \alpha(x) + \alpha(y). \quad (58)$$

Flood, in his PhD thesis [17] and a followup paper [18] called any pair  $(\rho, \alpha)$ , where  $\rho$  is a metric and  $\alpha$  a positive function, which satisfies the above property (58), a *normed pair*. The triple  $(X, \rho, \alpha)$ , where  $(\rho, \alpha)$  is a norm pair on  $X$ , is called a *normed set* [58]. Every normed space  $(E, \|\cdot\|_E)$  naturally becomes the normed set by setting  $\rho(x, y) = \|x - y\|_E$  and  $\alpha(x) = \|x\|_E$ .

For any two normed sets  $X_1 = (X_1, \rho_1, \alpha_1)$  and  $X_2 = (X_2, \rho_2, \alpha_2)$ , a function  $\pi : X_1 \rightarrow X_2$  is called a contraction if for all  $x \in X_1$ ,

$$\alpha_2(\pi(x)) \leq \alpha_1(x) \quad (59)$$

and for all  $x, y \in X_1$ ,

$$\rho_2(\pi(x), \pi(y)) \leq \rho_1(x, y). \quad (60)$$

According to a result of Flood [17, 18] (see also [58]), the normed pair structure supports a natural embedding of  $X$  into a Banach space with a certain universal property.

**Theorem 7.1** ([17, 18, 58]). *Let  $X = (X, \rho, \alpha)$  be a normed set. There exists a complete normed space  $B(X)$  and an embedding of  $X$  into  $B(X)$  as a normed subset such that every contraction  $\pi$  from  $X$  to a complete normed space  $E$  lifts to a unique linear contraction  $\bar{\pi}: B(X) \rightarrow E$ . The pair consisting of  $B(X)$  and embedding  $X \hookrightarrow B(X)$  is essentially unique. Elements of  $X$  are linearly independent.*  $\square$

Therefore, spaces of biological sequences with local similarity metric may be founded upon Banach (or even Hilbert) spaces. However, this result carries only theoretical significance at this point and cannot be directly used for clustering or indexing since the free Banach space  $B(X)$  is too large (it is not desirable that all sequences are linearly independent). Nevertheless, the same idea can be used to embed similarity score matrices into finite dimensional normed spaces and hence consider biological sequences as free semigroups over  $\mathbb{R}^n$ .

## Acknowledgments

A.S. is very grateful to Vladimir Pestov who as his Ph.D. and postdoctoral supervisor read and commented on the early versions of this manuscript. A.S. was supported by the University of Ottawa research funds. This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health.



## A Proofs

### A.1 General conditions for edit quasi-metrics

**Theorem A.1.** *Let  $\Sigma$  be a set, let  $1 \leq p < \infty$  and suppose  $d$  is a separating quasi-metric on  $\Sigma$ ,  $\alpha, \beta \in \Gamma$  and  $D$  is the  $\ell^p$  edit distance extending  $d, \alpha$  and  $\beta$ . In addition, assume that for all  $a, b \in \Sigma$ ,  $u, v, x \in \Sigma^*$ ,*

$$(W1) \quad d^p(a, b) + \beta^p(ubv) \geq \beta^p(uav);$$

$$(W2) \quad d^p(a, b) + \alpha^p(uav) \geq \alpha^p(ubv);$$

$$(W3) \quad \beta^p(uv) + \beta^p(x) \geq \beta^p(uxv);$$

$$(W4) \quad \alpha^p(uv) + \alpha^p(x) \geq \alpha^p(uxv);$$

$$(W5) \quad \beta^p(uxv) + \alpha^p(x) \geq \beta^p(uv);$$

$$(W6) \quad \alpha^p(uxv) + \beta^p(x) \geq \alpha^p(uv);$$

$$(W7) \quad \alpha^p(ux) + \beta^p(xv) \geq \alpha^p(u) + \beta^p(v);$$

$$(W8) \quad \beta^p(ux) + \alpha^p(xv) \geq \beta^p(u) + \alpha^p(v).$$

*Then,  $D$  is a separating quasi-metric on  $\Sigma^*$ .*

*Proof.* Let  $x, y, z \in \Sigma^*$ . Clearly,  $D(x, y)$  is non-negative since all of  $d$ ,  $\alpha$  and  $\beta$  are non-negative. Also,  $D(x, x) \leq (\sum_i d^p(x_i, x_i))^{1/p} = 0$ . Now suppose  $D(x, y) = 0$ . Applying Lemma 3.7, we have

$$D(x, y) = \left( \sum_{k=1}^K D^p(x_k^*, y_k^*) \right)^{1/p} = 0,$$

where  $x = x_1^* x_2^* \dots x_K^*$ ,  $y = y_1^* y_2^* \dots y_K^*$ , implying  $D(x_k^*, y_k^*) = 0$  for all  $k$  since  $D$  is non-negative. Hence,  $x_k^* = y_k^*$  for all possible cases of  $x_k^*$  and  $y_k^*$  because  $d$  is a separating quasi-metric and  $\alpha$  and  $\beta$  are strictly positive on  $\Sigma^+$ .

We will demonstrate the triangle inequality by relying on the Minkowski inequality: for any two sequences  $a$  and  $b$  of real numbers and  $1 \leq p < \infty$ ,

$$\left( \sum_i |a_i + b_i|^p \right)^{1/p} \leq \left( \sum_i |a_i|^p \right)^{1/p} + \left( \sum_i |b_i|^p \right)^{1/p}. \quad (61)$$

We show by induction that for all  $0 \leq i \leq |x|$ ,  $0 \leq j \leq |y|$  and  $0 \leq k \leq |z|$ ,

$$D(\bar{x}_i, \bar{y}_j) + D(\bar{y}_j, \bar{z}_k) \geq D(\bar{x}_i, \bar{z}_k). \quad (62)$$

Let  $\preceq$  denote a partial order on  $\mathbb{N} \times \mathbb{N} \times \mathbb{N}$  where  $(i_0, j_0, k_0) \preceq (i, j, k)$  if  $i_0 \leq i$  or  $i_0 = i$  and  $j_0 \leq j$  or  $i_0 = i$  and  $j_0 = j$  and  $k_0 \leq k$  (lexicographic order). The relation  $\preceq$  is a well-founded partial order of type  $\omega^3$  (in this case our induction is finite) and our claim is trivially true for  $(0, 0)$ . Assume it is true for all  $(i', j', k') \prec (i, j, k)$ . There are nine possibilities in total to consider for  $(i', j', k') = (i, j, k)$ .

**Case 1:** Suppose  $D(\bar{x}_i, \bar{y}_j) = (D^p(\bar{x}_{i-1}, \bar{y}_{j-1}) + d^p(x_i, y_j))^{1/p}$  and  $D(\bar{y}_j, \bar{z}_k) = (D^p(\bar{y}_{j-1}, \bar{z}_{k-1}) + d^p(y_j, z_k))^{1/p}$ . By the Minkowski inequality, our induction hypothesis and the triangle inequality on  $d$  we have

$$\begin{aligned} D(\bar{x}_i, \bar{y}_j) + D(\bar{y}_j, \bar{z}_k) &= (D^p(\bar{x}_{i-1}, \bar{y}_{j-1}) + d^p(x_i, y_j))^{1/p} \\ &\quad + (D^p(\bar{y}_{j-1}, \bar{z}_{k-1}) + d^p(y_j, z_k))^{1/p} \\ &\geq ((D(\bar{x}_{i-1}, \bar{y}_{j-1}) + D(\bar{y}_{j-1}, \bar{z}_{k-1}))^p \\ &\quad + (d(x_i, y_j) + d(y_j, z_k))^p)^{1/p} \\ &\geq (D^p(\bar{x}_{i-1}, \bar{z}_{k-1}) + d^p(x_i, z_k))^{1/p} \\ &\geq D(\bar{x}_i, \bar{z}_k). \end{aligned}$$

**Case 2:** Suppose  $D(\bar{y}_j, \bar{z}_k) = (D^p(\bar{y}_j, \bar{z}_{k-t}) + \alpha^p(z_{k-t+1} \dots z_k))^{1/p}$  for some  $1 \leq t < k$  (this covers three possibilities). By the Minkowski inequality and the induction hypothesis we have

$$\begin{aligned} D(\bar{x}_i, \bar{y}_j) + D(\bar{y}_j, \bar{z}_k) &= D(\bar{x}_i, \bar{y}_j) + (D^p(\bar{y}_j, \bar{z}_{k-t}) + \alpha^p(z_{k-t+1} \dots z_k))^{1/p} \\ &\geq ((D(\bar{x}_i, \bar{y}_j) + D(\bar{y}_j, \bar{z}_{k-t}))^p + \alpha^p(z_{k-t+1} \dots z_k))^{1/p} \\ &\geq (D^p(\bar{x}_i, \bar{z}_{k-t}) + \alpha^p(z_{k-t+1} \dots z_k))^{1/p} \\ &\geq D(\bar{x}_i, \bar{z}_k). \end{aligned}$$

**Case 3:** Suppose  $D(\bar{x}_i, \bar{y}_j) = (D^p(\bar{x}_{i-t}, \bar{y}_j) + \beta^p(x_{i-t+1} \dots x_i))^{1/p}$  for some  $1 \leq t < i$  (this covers additional two possibilities). Then, in similar manner as in Case 2,

$$D(\bar{x}_i, \bar{y}_j) + D(\bar{y}_j, \bar{z}_k) \geq (D^p(\bar{x}_{i-t}, \bar{z}_k) + \beta^p(x_{i-t+1} \dots x_i))^{1/p} \geq D(\bar{x}_i, \bar{z}_k),$$

by the Minkowski inequality and the induction hypothesis.

**Case 4:** Suppose  $D(\bar{y}_j, \bar{z}_k) = (D^p(\bar{y}_{j-t}, \bar{z}_k) + \beta^p(y_{j-t+1} \dots y_j))^{1/p}$ , for some  $1 \leq t < j$  (this covers additional two possibilities). Using Lemma 3.7, let  $0 \leq q \leq j$  be the smallest integer not larger than  $t$  such that

$$D(\bar{x}_i, \bar{y}_j) = \left( D^p(\bar{x}_r, \bar{y}_{j-q}) + \sum_{m=1}^K D^p(u_m^*, v_m^*) \right)^{1/p},$$

for some  $1 \leq r \leq i$ , where  $u = x_{r+1} \dots x_i = u_1^* \dots u_K^*$ , and  $v = y_{j-q+1} \dots y_j = v_1^* \dots v_K^*$ . Note that  $q < t$  if and only if  $D(\bar{x}_r, \bar{y}_{j-q}) = (D^p(\bar{x}_r, \bar{y}_{j-q'}) + D^p(e, y_{j-q'+1} \dots y_{j-q}))^{1/p}$ , where  $q < t < q' \leq j$ . In that case, by our assumption (W7) and by Minkowski inequality,

$$\begin{aligned} D^p(\bar{x}_r, \bar{y}_{j-q}) + \beta^p(y_{j-t+1} \dots y_j) &= D^p(\bar{x}_r, \bar{y}_{j-q'}) + \alpha^p(y_{j-q'+1} \dots y_{j-q}) \\ &\quad + \beta^p(y_{j-t+1} \dots y_j) \\ &\geq D^p(\bar{x}_r, \bar{y}_{j-q'}) + \alpha^p(y_{j-q'+1} \dots y_{j-t}) \\ &\quad + \beta^p(y_{j-q+1} \dots y_j) \\ &\geq D^p(\bar{x}_r, \bar{y}_{j-t}) + \beta^p(v). \end{aligned}$$

Of course, the same inequality trivially holds if  $t = q$ .

Observe that assumptions (W1), (W3) and (W5) imply that for any  $1 \leq m \leq K$  and any  $w_1, w_2 \in \Sigma^*$ ,

$$D^p(u_m^*, v_m^*) + \beta^p(w_1 v_m^* w_2) \geq \beta^p(w_1 u_m^* w_2), \quad (63)$$

and hence

$$\begin{aligned} \sum_{m=1}^K D^p(u_m^*, v_m^*) + \beta^p(y_{j-q+1} \dots y_j) &\geq \sum_{m=2}^K D^p(u_m^*, v_m^*) + \beta^p(u_1^* v_2^* \dots v_K^*) \\ &\geq \sum_{m=3}^K D^p(u_m^*, v_m^*) + \beta^p(u_1^* u_2^* v_3^* \dots v_K^*) \\ &\geq \beta^p(u_1^* \dots u_K^*) \\ &= \beta^p(u). \end{aligned}$$

Therefore,

$$\begin{aligned}
D(\bar{x}_i, \bar{y}_j) + D(\bar{y}_j, \bar{z}_k) &= \left( D^p(\bar{x}_r, \bar{y}_{j-q}) + \sum_{m=1}^K D^p(u_m^*, v_m^*) \right)^{1/p} \\
&\quad + \left( D^p(\bar{y}_{j-t}, \bar{z}_k) + \beta^p(y_{j-t+1} \dots y_j) \right)^{1/p} \\
&\geq \left( D^p(\bar{x}_r, \bar{y}_{j-q}) + \sum_{m=1}^K D^p(u_m^*, v_m^*) + D^p(\bar{y}_{j-t}, \bar{z}_k) \right. \\
&\quad \left. + \beta^p(y_{j-t+1} \dots y_j) \right)^{1/p} \\
&\geq \left( D^p(\bar{x}_r, \bar{y}_{j-t}) + \sum_{m=1}^K D^p(u_m^*, v_m^*) + D^p(\bar{y}_{j-t}, \bar{z}_k) \right. \\
&\quad \left. + \beta^p(y_{j-q+1} \dots y_j) \right)^{1/p} \\
&\geq \left( D^p(\bar{x}_r, \bar{z}_k) + \beta^p(u) \right)^{1/p} \\
&\geq D(\bar{x}_i, \bar{z}_k),
\end{aligned}$$

by the induction hypothesis.

**Case 5:** The remaining case is  $D(\bar{x}_i, \bar{y}_j) = (D^p(\bar{x}_i, \bar{y}_{j-t}) + \alpha^p(y_{j-t+1} \dots y_j))^{1/p}$  for some  $1 \leq t < j$  and  $D(\bar{y}_j, \bar{z}_k) = (D^p(\bar{y}_{j-1}, \bar{z}_{k-1}) + d^p(y_j, z_k))^{1/p}$ . The proof for this case exactly mirrors the proof for the previous case, now depending on the assumptions (W2), (W4), (W6) and (W8).  $\square$

**Remark A.2.** In general the assumptions (W1) – (W8) are sufficient for  $D$  to be a quasi-metric but not necessary, except in the case of  $p = 1$ . For example, let  $\Sigma = \{a, b\}$ ,  $d(a, b) = d(b, a) = 3$ ,  $\alpha = \beta$ ,  $\alpha(a) = 7$ ,  $\alpha(b) = 4$ ,  $\alpha(u) = \sum_i \alpha(u_i)$ . In this case the assumptions (W1) and (W2) fail but it can be verified that the triangle inequality for  $D$  does not fail for any  $p > 1$ .

**Remark A.3.** The assumptions (W1)–(W8) can be significantly simplified if the gap penalties take a more restricted form. For example, if the gap penalties are increasing, the assumptions (W5)–(W8) can be removed. This restriction is sensible in applications to biological sequence comparisons because algebraic interactions lowering the effective length of the sequence are

not allowed. On the other hand, if  $\Sigma^*$  is replaced as the underlying set with a monoid which is not free, or even a group, then gap penalties cannot be increasing in the above sense.

Since composition-length gap penalties are increasing by definition, Theorem 3.10 is a direct corollary of Theorem A.1. Furthermore, composition-length gap penalties with  $\phi = 0$ , such as linear or affine, satisfy all of (W1)–(W8).

## A.2 Global similarities

**Proposition 4.5.** *Let  $\Sigma$  be a set and let  $s : \Sigma \times \Sigma \rightarrow \mathbb{R}$  be a sane scoring function over  $\Sigma$ . Suppose  $\gamma, \delta \in \Gamma(\Sigma)$  and  $S$  the global similarity on  $\Sigma^*$  with respect to  $s, \delta$  and  $\gamma$ . Then,  $S$  is a sane scoring function and for all  $x \in \Sigma^*$ ,*

$$S(x, x) = \sum_{i=1}^{|x|} s(x_i, x_i). \quad (64)$$

□

We will make use of the following lemma, equivalent to Lemma 3.7 for distances. It was likewise proved by Smith and Waterman [68] for the  $\ell^1$  case and less general gap penalties.

**Lemma A.4.** *Let  $\Sigma$  be a set,  $s : \Sigma \times \Sigma \rightarrow \mathbb{R}$ , and  $\gamma, \delta : \Sigma^+ \rightarrow \mathbb{R}_+$ . Suppose  $S$  is a global similarity on  $\Sigma^*$  with respect to  $d, \gamma$  and  $\delta$ . Then, for all  $x, y \in \Sigma^*$*

$$S(x, y) = \max \left\{ \sum_{k=1}^K S(x_k^*, y_k^*) \mid \langle (x_k^*, y_k^*) \rangle_{k=1}^K \in \mathcal{A}(x, y) \right\}. \quad (65)$$

*Proof of Proposition 4.5.* Let  $x, y \in \Sigma^*$ . If  $x = e$ , by definition  $S(x, x) = 0$ , coinciding with a sum over the empty set. Since  $\gamma$  and  $\delta$  are positive, we have  $-\gamma(y) = S(e, y) \leq 0$  and  $-\delta(y) = S(y, e) \leq 0$ .

Now suppose  $x \in \Sigma^+$  and let  $\langle (x_k^*, y_k^*) \rangle_{k=1}^K \in \mathcal{A}(x, y)$  such that  $S(x, y) = \sum_{k=1}^K S(x_k^*, y_k^*)$ . Let  $C = \{k : x_k^* \in \Sigma \text{ and } y_k^* \in \Sigma\}$  and  $D = \{k : x_k^* \in$

$\Sigma^+$  and  $y_k^* = e\}$ . Then,

$$\begin{aligned}
S(x, y) &\leq \sum_{k \in C} S(x_k^*, y_k^*) + \sum_{k \in D} S(x_k^*, y_k^*) \\
&\leq \sum_{k \in C} s(x_k^*, y_k^*) - \sum_{k \in D} \delta(x_k^*) \\
&\leq \sum_{k \in C} s(x_k^*, x_k^*) + \sum_{k \in D} \sum_j s((x_k^*)_j, (x_k^*)_j) \\
&= \sum_{i=1}^{|x|} s(x_i, x_i),
\end{aligned}$$

since  $s$  is sane and the whole of  $x$  is accounted for in fragments indexed by  $C$  and  $D$ . Therefore,

$$S(x, y) \leq \sum_{i=1}^{|x|} s(x_i, x_i) \leq S(x, x), \quad (66)$$

implying  $S(x, x) = \sum_{i=1}^{|x|} s(x_i, x_i) > 0$  and  $S(x, x) \geq S(x, y)$ . In the same way it can be shown that  $S(x, x) \geq S(y, x)$  and hence that  $S$  is sane.  $\square$

**Corollary 4.7.** *Let  $\Sigma$  be a set and let  $1 \leq p < \infty$ . Suppose  $s$  is a sane scoring function on  $\Sigma$ ,  $d = \mathbf{AQ}^p(s)$  is a quasi-metric on  $\Sigma$  and  $\gamma, \delta \in \Gamma_{CL}(\Sigma)$  such that*

$$\gamma(b) - \gamma(a) \leq d^p(a, b) \quad (67)$$

and

$$s(a, a) + \delta(a) - s(b, b) - \delta(b) \leq d^p(a, b). \quad (68)$$

Let  $S$  be the global similarity with respect to  $s, \gamma$  and  $\delta$  and let  $\alpha(x) = \gamma(x)^{1/p}$  and  $\beta(x) = (S(x, x) + \delta(x))^{1/p}$  for all  $x \in \Sigma^+$ . Then, the  $\ell^p$  edit distance  $D = \mathbf{EQ}^p(d, \alpha, \beta)$  is given for all  $x, y \in \Sigma^*$  by the formula

$$D(x, y) = \left( S(x, x) - S(x, y) \right)^{1/p}. \quad (69)$$

*Proof.* By construction,  $\alpha^p \in \Gamma_{CL}(\Sigma)$  and by Proposition 4.5,  $\beta^p \in \Gamma_{CL}(\Sigma)$  as well. By our assumptions on  $d, \gamma$  and  $\delta$  and by Theorem 3.10, it follows that  $D$ , the  $\ell^p$  edit distance extending  $d, \alpha$  and  $\beta$ , is indeed the separating

quasi-metric  $\mathbf{EQ}^p(d, \alpha, \beta)$  on  $\Sigma^*$ . We will now show by recursion that this quasi-metric is equivalent to the one given by Equation (69).

Clearly,  $D(e, e) = (S(e, e) - S(e, e))^{1/p} = 0$ . Let  $x, y \in \Sigma^+$  and suppose  $1 \leq i \leq |x|$  and  $1 \leq j \leq |y|$ . We have,

$$D(e, \bar{y}_j) = \alpha(\bar{y}_j) = \gamma(\bar{y}_j)^{1/p} = (S(e, e) - S(e, \bar{y}_j))^{1/p},$$

and

$$D(\bar{x}_i, e) = \beta(\bar{x}_i) = (\delta(\bar{x}_i) + S(\bar{x}_i, \bar{x}_i))^{1/p} = (S(\bar{x}_i, \bar{x}_i) - S(\bar{x}_i, e))^{1/p}.$$

Using recursion and Proposition 4.5,

$$\begin{aligned} D(\bar{x}_i, \bar{y}_j) &= \left( \min \left\{ D^p(\bar{x}_{i-1}, \bar{y}_{j-1}) + d^p(x_i, y_j), \right. \right. \\ &\quad \min_{1 \leq k \leq j} \{ D^p(\bar{x}_i, \bar{y}_{j-k}) + \alpha^p(y_{j-k+1} \dots y_j) \}, \\ &\quad \left. \left. \min_{1 \leq k \leq i} \{ D^p(\bar{x}_{i-k}, \bar{y}_j) + \beta^p(x_{i-k+1} \dots x_i) \} \right\} \right)^{1/p} \\ &= \left( \min \left\{ S(\bar{x}_{i-1}, \bar{x}_{i-1}) - S(\bar{x}_{i-1}, \bar{y}_{j-1}) + s(x_i, x_i) - s(x_i, y_j), \right. \right. \\ &\quad \min_{1 \leq k \leq j} \{ S(\bar{x}_i, \bar{x}_i) - S(\bar{x}_i, \bar{y}_{j-k}) + \gamma(y_{j-k+1} \dots y_j) \}, \\ &\quad \min_{1 \leq k \leq i} \{ S(\bar{x}_{i-k}, \bar{x}_{i-k}) - S(\bar{x}_{i-k}, \bar{y}_j) + \delta(x_{i-k+1} \dots x_i) \\ &\quad \left. \left. + S(x_{i-k+1} \dots x_i, x_{i-k+1} \dots x_i) \} \right\} \right)^{1/p} \\ &= \left( S(\bar{x}_i, \bar{x}_i) - \max \left\{ S(\bar{x}_{i-1}, \bar{y}_{j-1}) + s(x_i, y_j), \right. \right. \\ &\quad \max_{1 \leq k \leq j} \{ S(\bar{x}_i, \bar{y}_{j-k}) - \gamma(y_{j-k+1} \dots y_j) \}, \\ &\quad \left. \left. \max_{1 \leq k \leq i} \{ S(\bar{x}_{i-k}, \bar{y}_j) - \delta(x_{i-k+1} \dots x_i) \} \right\} \right)^{1/p} \\ &= \left( S(\bar{x}_i, \bar{x}_i) - S(\bar{x}_i, \bar{y}_j) \right)^{1/p}, \end{aligned}$$

as required.  $\square$

## References

- [1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.
- [2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
- [3] S. F. Altschul, J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schaffer, and Y.-K. Yu. Protein database searches using compositionally adjusted substitution matrices. *FEBS J.*, 272(20):5101–5109, 2005.
- [4] R. Bellman, J. Holland, and R. Kalaba. On an application of dynamic programming to the synthesis of logical systems. *J. ACM*, 6(4):486–493, 1959.
- [5] J. D. Blake and F. E. Cohen. Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.*, 307(2):721–735, Mar 2001.
- [6] J. Buhler. Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, 17:419–428, 2001.
- [7] G. E. Crooks and S. E. Brenner. An alternative model of amino acid replacement. *Bioinformatics*, 21(7):975–980, 2005.
- [8] S. Dasgupta and P. M. Long. Performance guarantees for hierarchical clustering. *J. Comput. Syst. Sci.*, 70(4):555–569, 2005.
- [9] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, chapter 22, pages 345–352. National Biomedical Research Foundation, 1978.
- [10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge University press, Cambridge, UK, 1998.
- [11] S. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.



- [12] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.
- [13] R. C. Edgar and K. Sjölander. A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, 20(8):1301–1308, 2004.
- [14] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE T. Inform. Theory*, 49(7):1858–1860, 2003.
- [15] D. F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25(4):351–360, 1987.
- [16] I. Fischer. Similarity-preserving metrics for amino-acid sequences. Poster at the 22nd GIF Meeting on Challenges in Genomic Research: Neurodegenerative Diseases, Stem Cells, Bioethics, Heidelberg 2002.
- [17] J. Flood. *Free Topological Vector Spaces*. PhD thesis, Australian National University, Canberra, 1975. 109 pp.
- [18] J. Flood. Free topological vector spaces. *Dissertationes Math. (Rozprawy Mat.)*, 221:95 pp., 1984.
- [19] E. Giladi, M. G. Walker, J. Z. Wang, and W. Volkmuth. SST: an algorithm for finding near-exact sequence matches in time proportional to the logarithm of the database size. *Bioinformatics*, 18(6):873–877, 2002.
- [20] G. Gonnet, M. Cohen, and S. Benner. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445, 1992.
- [21] O. Gotoh. An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162:705–708, 1982.
- [22] M. I. Graev. Free topological groups. *Izvestiya Akad. Nauk SSSR. Ser. Mat.*, 12:279–324, 1948.
- [23] M. I. Graev. Free topological groups. *Amer. Math. Soc. Translation*, 1951(35):61, 1951.

- [24] M. Gribskov, A. D. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 84:4355–4358, 1987.
- [25] D. Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.
- [26] R. W. Hamming. Error detecting and error correcting codes. *Bell System Tech. J.*, 29:147–160, 1950.
- [27] J. M. Hellerstein, E. Koutsoupias, and C. H. Papadimitriou. On the analysis of indexing schemes. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS’97) (Tucson, Arizona, May)*, pages 249–256, 1997.
- [28] S. Henikoff and J. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89:10915–10919, 1992.
- [29] S. Henikoff and J. G. Henikoff. Performance evaluation of amino acid substitution matrices. *Proteins*, 17(1):49–61, 1993.
- [30] G. R. Hjaltason and H. Samet. Index-driven similarity search in metric spaces. *ACM Trans. Database Syst.*, 28(4):517–580, 2003.
- [31] E. Hunt. Indexed Searching on Proteins Using a Suffix Sequoia. *IEEE Data Eng. Bull.*, 27:24–31, 2004.
- [32] E. Hunt, M. P. Atkinson, and R. W. Irving. A database index to large biological sequences. *VLDB J.*, 11(3):139–148, 2001.
- [33] M. Itoh, S. Goto, T. Akutsu, and M. Kanehisa. Fast and accurate database homology search using upper bounds of local alignment scores. *Bioinformatics*, 21(7):912–921, 2005.
- [34] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8(3):275–282, 1992.
- [35] D. T. Jones, W. R. Taylor, and J. M. Thornton. A mutation data matrix for transmembrane proteins. *FEBS Lett.*, 339(3):269–275, Feb 1994.

- [36] M. Kann, B. Qian, and R. A. Goldstein. Optimization of a new score function for the detection of remote homologs. *Proteins*, 41(4):498–503, Dec 2000.
- [37] S. Karlin and S. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 90(12):5873–5877, 1993.
- [38] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.*, 87:2264–2268, 1990.
- [39] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30(14):3059–3066, Jul 2002.
- [40] S. Kawashima, H. Ogata, and M. Kanehisa. AAindex: amino acid index database. *Nucleic Acids Res.*, 27:368–369, 1999.
- [41] W. J. Kent. BLAT—the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664, 2002.
- [42] M. Kschischo, M. Lssig, and Y. Yu. Toward an accurate statistics of gapped alignments. *Bull. Math. Biol.*, 67(1):169–91, 2005.
- [43] H.-P. A. Küenzi. Nonsymmetric distances and their associated topologies: about the origins of basic ideas in the area of asymmetric topology. In *Handbook of the history of general topology, Vol. 3*, volume 3 of *Hist. Topol.*, pages 853–968. Kluwer Acad. Publ., Dordrecht, 2001.
- [44] H.-P. A. Küenzi and V. Vajner. Weighted quasi-metrics. In *Papers on general topology and applications (Flushing, NY, 1992)*, pages 64–77. New York Acad. Sci., New York, 1994.
- [45] T. Lassmann and E. L. L. Sonnhammer. Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6:298, 2005.
- [46] V. I. Levenstein. Binary codes capable of correcting insertions and reversals. *Sov. Phys. Dokl.*, pages 707–710, 1966.

- [47] M. Linial, N. Linial, N. Tishby, and G. Yona. Global self organization of all known protein sequences reveals inherent biological signatures. *J. Mol. Biol.*, 268:539–556, 1997.
- [48] R. Mao, W. Xu, N. Singh, and D. P. Miranker. An assessment of a metric space database index to support sequence homology. In *3rd IEEE International Symposium on BioInformatics and BioEngineering (BIBE 2003)*, (Bethesda, Maryland, March 2003), pages 375–384, 2003.
- [49] M. A. Marti-Renom, M. Madhusudhan, and A. Sali. Alignment of protein sequences by their profiles. *Protein Sci.*, 13(4):1071–1087, 2004.
- [50] S. G. Matthews. Partial metric topology. In *Papers on general topology and applications (Flushing, NY, 1992)*, volume 728 of *Ann. New York Acad. Sci.*, pages 183–197. New York Acad. Sci., New York, 1994.
- [51] T. Müller, S. Rahmann, and M. Rehmsmeier. Non-symmetric score matrices and the detection of homologous transmembrane proteins. In *ISMB (Supplement of Bioinformatics)*, pages 182–189, 2001.
- [52] T. Müller, R. Spang, and M. Vingron. Estimating Amino Acid Substitution Models: A Comparison of Dayhoff’s Estimator, the Resolvent Approach and a Maximum Likelihood Method. *Mol. Biol. Evol.*, 19(1):8–13, 2002.
- [53] T. Müller and M. Vingron. Modeling amino acid replacement. *J. Comput. Biol.*, 7(6):761–776, 2000.
- [54] K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.*, 2:93–100, 1988.
- [55] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.
- [56] P. C. Ng, J. G. Henikoff, and S. Henikoff. PHAT: a transmembrane-specific substitution matrix. *Bioinformatics*, 16(9):760–766, 2000.
- [57] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 85:2444–2448, 1988.

- [58] V. Pestov. Douady’s conjecture on Banach analytic spaces. *C. R. Acad. Sci. Paris Sér. I Math.*, 319(10):1043–1048, 1994.
- [59] V. Pestov. Topological groups: where to from here? *Topology Proc.*, 24:421–502, 1999.
- [60] V. Pestov. On the geometry of similarity search: dimensionality curse and concentration of measure. *Inform. Process. Lett.*, 73:47–51, 2000.
- [61] V. Pestov and A. Stojmirović. Indexing schemes for similarity search: an illustrated paradigm. *Fundam. Inform.*, 70(4):367–385, 2006.
- [62] S. Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments [published erratum appears in *Nucleic Acids Res* 1996 Nov 1;24(21):4372]. *Nucl. Acids Res.*, 24(19):3836–3845, 1996.
- [63] A. Prlic, F. S. Domingues, and M. J. Sippl. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.*, 13(8):545–550, 2000.
- [64] S. Romaguera and M. P. Schellekens. Weightable quasi-metric semi-groups and semilattices. *Electr. Notes Theor. Comput. Sci.*, 40, 2000.
- [65] L. Rychlewski, L. Jaroszewski, W. Li, and A. Godzik. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, 9(2):232–241, 2000.
- [66] O. Sasson, N. Linial, and M. Linial. The metric space of proteins—comparative study of clustering algorithms. *Bioinformatics*, 18(suppl.1):S14–21, 2002.
- [67] P. H. Sellers. On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, 26:787–793, 1974.
- [68] T. F. Smith and M. S. Waterman. Comparison of biosequences. *Adv. in Appl. Math.*, 2(4):482–489, 1981.
- [69] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

- [70] T. F. Smith, M. S. Waterman, and W. M. Fitch. Comparative biosequence metrics. *J. Mol. Evol.*, 18:38–46, 1981.
- [71] P. A. Spiro and N. Macura. A local alignment metric for accelerating biosequence database search. *J. Comput. Biol.*, 11(1):61–82, 2004.
- [72] D. J. States, W. Gish, and S. F. Altschul. Improved sensitivity of nucleic acid database similarity searches using application specific scoring matrices. *Methods: A companion to Methods in Enzymology*, 3:66–70, 1991.
- [73] A. Stojmirović. Quasi-metric spaces with measure. *Topology Proc.*, 28(2):655–671, 2004.
- [74] A. Stojmirović and V. Pestov. Indexing schemes for similarity search in datasets of short protein fragments. *Inf. Syst.*, 32(8):1145–1165, 2007.
- [75] Z. Tan, X. Cao, B. C. Ooi, and A. K. H. Tung. The ed-tree: an index for large dna sequence databases. In *SSDBM’2003: Proceedings of the 15th international conference on Scientific and statistical database management*, pages 151–160, Washington, DC, USA, 2003. IEEE Computer Society.
- [76] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, November 1994.
- [77] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, 9:27–36, 1996.
- [78] S. Veerassamy, A. Smith, and E. R. M. Tillier. A transition probability model for amino acid substitutions from blocks. *J. Comput. Biol.*, 10(6):997–1010, 2003.
- [79] P. Vitolo. The representation of weighted quasi-metric spaces. *Rend. Istit. Mat. Univ. Trieste*, 31(1-2):95–100, 1999.
- [80] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J. Comput. Biol.*, 1(4):337–348, 1994.

- [81] M. S. Waterman. Efficient sequence alignment algorithms. *J. Theor. Biol.*, 108(3):333–337, 1984.
- [82] M. S. Waterman, T. F. Smith, and W. A. Beyer. Some biological sequence metrics. *Advances in Math.*, 20(3):367–387, 1976.
- [83] W. Wu, H. Xiong, and S. Shekhar, editors. *Clustering and Information Retrieval*. Kluwer, 2003.
- [84] G. Yona and M. Levitt. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, 315(5):1257–1275, 2002.
- [85] Y.-K. Yu. A metric measure for weight matrices of variable lengths – with applications to clustering and classification of hidden Markov models. *Physica A*, 375:212–220, 2007.
- [86] Y.-K. Yu and S. F. Altschul. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, 21(7):902–911, Apr 2005.
- [87] Y.-K. Yu, J. C. Wootton, and S. F. Altschul. The compositional adjustment of amino acid substitution matrices. *Proc. Natl. Acad. Sci. U.S.A.*, 100(26):15688–15693, 2003.